A Holistic Design Concept For Eyes-Free Mobile Interfaces

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Doctor of Philosophy
in the
University of Canterbury
by
Christina Dicke

Examining Committee

Tim Bell	Supervisor
Mark Billinghurst	Supervisor

University of Canterbury
2012

This thesis is dedicated to my mother, who has always given me courage to pursue my dreams.

Declaration

The material presented in this thesis is the result of my own research carried out at the Department of Computer Science at the University of Canterbury and other institutions working under the supervision of Professor Tim Bell and Professor Mark Billinghurst.

- The work contained in chapter 2 has been partly published as [1].
- The work contained in chapter 3 has been carried out at the Hasso Plattner Institut in Potsdam, Germany in cooperation with Patrick Baudisch.
- The work contained in chapter 6 has been carried out at the Nokia Research Center in Helsinki, Finland, and has been published as [2].
- The work contained in chapter 5 has been carried out at the Nokia Research Center in Helsinki, Finland, and has been published as [3, 4].
- The work contained in chapter 4 has been carried out at the Aalto University in Aalto, Finland in cooperation with Kai Puolamaki.
- The work contained in chapter 7 has been carried out at the Human Interface Technology Laboratory, Christchurch, New Zealand. Section 7.1 has been published as [5], section 7.2 has been published as [6] and section 7.3 has been published as [7].
- The work contained in chapter 8 has been carried out at the Human Interface Technology Laboratory, Christchurch, New Zealand, and has been published as [8].

Abstract

This thesis presents a series of studies to explore and understand the design of eyes-free interfaces for mobile devices. The motivation is to devise a holistic design concept that is based on the WIMP paradigm and is adapted to the requirements of mobile user interaction. It is proposed that audio is a very efficient and effective modality for use in an eyes-free mobile interface. Methods to transfer the WIMP paradigm to eyes-free interfaces are proposed and evaluated. Guidelines for the implementation of the paradigm are given and – by means of an example – a holistic design concept is proposed.

This thesis begins with an introduction to and critical reflection of recurrently important themes and research methods from the disciplines of psychoacoustics, psychology, and presence research. An overview of related work is given, paying particular attention to the use of interface metaphors in mobile eyes-free interfaces. The notion of distance is discussed as a method to prioritise, structure, and manage attention in eyes-free interfaces. Practical issues arising from sources becoming inaudible with increasing distance can be addressed by proposing a method modeled on echo location. This method was compared to verbally coded distance information and proved useful for identifying the closest of several objects, while verbally coded distance information was found to be more efficient for identifying the precise distance of an object. The knowledge gained from the study can contribute to improving other applications, such as GPS based navigation. Furthermore, the issue of gaining an overview of accessible objects by means of sound was examined. The results showed that a minimum of 200 ms between adjacent sound samples should be adhered to. Based on these findings, both earcons and synthesized speech are recommendable, although speech has the advantage of being more flexible and easier to learn. Monophonic reproduction yields comparable results to spatial reproduction. However, spatial reproduction has the additional benefit of indicating an item's position. These results are transferable and generally relevant for the use of audio in HCI.

Tactile interaction techniques were explored as a means to interact with an auditory interface and were found to be both effective and enjoyable. One of the more general observations was that 2D and 3D gestures were intuitively used by participants, who transferred their knowledge of established gestures to auditory interfaces. It was also found that participants often used 2D gestures to select an item and proceeded to manipulate it with a 3D gesture. The results suggest the use of a small gesture set with reversible gestures for do/undo-type actions, which was further explored in a follow up study. It could be shown that simple 3D gestures are a viable way of manipulating spatialized sound sources in a complex 3D auditory display.

While the main contribution of this thesis lies in the area of HCI, previously unresearched issues from adjacent disciplines that impact the user experience of auditory interfaces have been addressed. It was found that regular, predictable movement patterns in 3D audio spaces cause symptoms of simulator sickness. However, these were found to be minor and only occurred under extreme conditions. Additionally, the influence of the audio reproduction method on the perception of presence, social presence, and realism was examined. It was found that both stereophonic and binaural reproduction have advantages over monophonic sound reproduction: stereophonic sound increases the perception of social presence while binaural sound increases the feeling of being present in a virtual environment. The results are important contributions insofar as one of the main applications of mobile devices is voice based communication; it is reasonable to assume that there will be an increase in real-time voice based social and cooperative networking applications.

This thesis concludes with a conceptual design of a system called “Foogue”, which uses the results of the previous experiments as the basis of an eyes-free interface that utilizes spatial audio and gesture input.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AR	Augmented Reality
ARA	Augmented Reality Audio
DBAB	Distance Based Amplitude Panning
DC	Direct current
CB	Critical Band (Def.: Frequency bandwidth of the auditory filter; Each band corresponds with a section within the cochlea.)
CLT	Cognitive Load Theory
CPA	Continuous Partial Attention
CSCW	Computer Supported Cooperative Work
dB	Decibel (Def.: Logarithmic unit that indicates the ratio of a physical quantity relative to a specified or implied reference level; Quantification of sound pressure levels (see **SPL**) relative to a 0 dB reference, which has been defined as typical threshold of perception of an average human.)
DOF	Degrees of freedom (Def.: The set of independent displacements and/or rotations that specify the displaced or deformed position and orientation of the system.)
EEG	Electroencephalography
ETA	Electronic travel aid
GPS	Global Positioning System

GUI	Graphical User Interface
HCI	Human-Computer Interaction
HDS	High Density Sonification (Def.: The simultaneous sonification of all or some data points in a set.)
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
Hz	Hertz (Def.: Number of cycles per second of a periodic phenomenon such as a sine wave. Humans perceive frequency of sound waves as pitch. Each musical note corresponds to a particular frequency, which can be measured in hertz.)
ILD	Inter Aural Level Difference
ITD	Inter Aural Time Difference
ISOI	Interstimulus Onset Interval
jnd	Just Noticeable Difference (The ear's resolving power for simultaneous tones or partials as detected on 50 percent of occasions by a particular response.)
kHz	Kilohertz
m	Meter
MAA	Minimum Audible Angle (Def.: The just distinguishable horizontal angular deviation between two sound sources.)
MAMA	Minimum Audible Movement Angle (Def.: The minimum angle of movement required for detection of the direction of sound movement.)

ms	Millisecond
μPA	Micropascal (1 μPa = 1/1,000,000 Pa)
ORTF	Office de Radiodiffusion Télévision Française (Def.: Microphone recording technique for stereo sound using two cardioid microphones spread to a 110 degrees angle.)
Pa	Pascal (Def.: Force per unit area; defined as one newton per square metre.)
PC	Personal Computer
PDA	Personal Digital Assistant
PPAM	Pre-perceptual Auditory Memory
RF	Radio Frequency
STAM	Post-perceptual Short-term Auditory Memory
QUIS	Questionnaire for User Interaction Satisfaction
sec	Second
SIM	Sound Instant Message
snr or s/n	Signal-to-Noise Ratio
SPL	Sound Pressure Level (Def.: A logarithmic measure of the sound pressure of a sound relative to a reference value. Measured in dB above a standard reference level (20 μPa, 0 dB or the threshold of hearing))
SSQ	Simulator Sickness Questionnaire
tts	Text-To-Speech

CBAP Vector Based Amplitude Panning

WIMP Windows, Icons, Menu, Pointing Device Paradigm

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Tim Bell, for his help, enthusiasm, encouragement, and invaluable advice over the course of this research project. I would also like to thank my second supervisor, Professor Mark Billinghurst, for bringing me to New Zealand, his sound advice, and inspiration. Thanks are also due to Professor Andy Cockburn and the members of the HCI research group at the University of Canterbury. I am also grateful to my colleagues from the Human Interface Technology Laboratory NZ. In particular, I want to thank Dr. Andreas Dünser, who talked me through most of my statistical problems, Dr. Julian Looser, who always found the missing '}' in my code, and Dr. Raphael Grasset, for his dedication, energy, and the many interesting discussions. I am truly indebted and thankful for the funding provided by the HIT Lab and the support from its staff, in particular Ken Beckman and Katy Bang. I also wish to express my warm and sincere thanks to Professor Thomas Furness III for reminding me to think outside the box.

Additionally, I want to thank Professor Hartmut Wandke at the Humboldt University in Berlin for encouraging me to start working on a Ph.D. I would also like to thank the members of the Immersive Communication Team at the Nokia Research Center in Helsinki and Tampere, Finland. Special thanks goes to Dr. Martin Schrader, Viljakaisa Aaltonen (especially, for making me jump into an almost frozen lake after the sauna by claiming it is a 'tradition'), Anssi Rämö and Miika Vilermo for their expertise in sound engineering, Dr. Johan Kildal for his feedback on my research, and Dr. Akos Vetek for being a truly mad scientist (this is meant as a compliment). I am very grateful for the opportunity to work with and learn from Professor Patrick Baudisch and his research students Sean Gustafson and Christian Holz at the Hasso Plattner Institut, University of Potsdam, Germany. Many thanks go also to Professor Kai Puolamaki and his research student Hannes Gamper who I worked with at the Aalto University, Finland. I would also

like to thank my colleagues Dr. Jaka Sodnik at the University of Ljubljana, Slovenia and Katrin Wolf from the Telekom Laboratories at the Technical University, Berlin.

Particular thanks go to the students and colleagues who volunteered for my experimental studies. I would like to dearly apologize to the 82 participants from Tampere who suffered through 2 hours of simulator sickness inducing experiments. It was for the sake of science!

Last but not least, I cannot overstate the importance of the support received from the people that have been at my side during the last four years. My biggest thanks go to Stefanie Burgert, Gordon Love, Yaroslav Tal, Dr. Kali Tal, and Heidi Vanttaja for their friendship, support, and care. I am also hugely grateful to my friend and mentor Dr. Mechthilde Vahsen, who always takes the time to explain the world to me when I call her. My mother, Urusla Haarmann, deserves a special mention for letting her only daughter move to the other side of the world without complaint, for encouraging me, caring for me, and being the best mother there is.

Chapter I

Introduction

About one year before I started working on this dissertation I was strolling through the Tiergarten, a large park in the heart of Berlin. I was heading for a café at the other side of the park to meet with a friend. It was spring, the weather was fine and I was quite enjoying myself. Then I received a text message. My mobile phone beeped, I took it out of my bag, read the message, and replied. After that I put it back in my bag. A few moments later it beeped again and I repeated the procedure once more.

That is when I noticed that all my attention had shifted away from enjoying the park towards typing and reading messages on the tiny screen of my mobile phone. I had been a designer for visual interfaces before but that day I became interested in alternative ways to present information and the design of human-computer interfaces that are adapted to the challenges posed when such devices are used in mobile contexts.

After researching the subject I learned that despite the inadequacy of the interfaces on the current generation of mobile phones [10, 11] (leading to bans being placed upon mobile phone use while driving in some countries), better interface concepts have not been realised, or at least they are nowhere to be *seen*.

The dominance of the WIMP paradigm made much more sense to me only after reading Lakoff & Johnson's [12] work on how our everyday perception of the world is formed around conceptual metaphors and understanding that we often think of abstract concepts in terms of objects, like *chewing on a problem*, or in terms of directionality, when *our mood has risen*[1].

A simple example of conceptual metaphors used in human-computer interaction illustrated by the way in which we think of *files* as coherent pieces

[1] Section 2.6.4 of chapter 2 is dedicated to a discussion of Lakoff's & Johnson's work in the context of interface metaphors for auditory displays.

of information. They can be moved or copied or erased, and are contained in folders, which in turn can be opened or closed and so forth. This is how we are used to thinking about data on our computers. However, when we refer to the representation of a file, we usually refer to fragmented, binary representations of pieces of information stored in different patterns of magnetization on a magnetically coated surface[2].

Acknowledging that metaphorical thinking about computers that goes way beyond simple interface metaphors such as the *desktop* is already established and resembles how we, as embodied beings, make use of our sensorimotor systems to interact with the world around us, the success of the WIMP paradigm can be explained by the way it resolves the physical gap between us and the machine:

Windows are conceptual frames giving access to, but also containing, applications or *views* on data, while *icons* help us to grasp the essence of what is referenced visually: a recycling bin signifies the option to delete objects and an envelope classifies a file to be of the type 'email'. Icons allow us to access what is signified as a *unit*.

Menus grant access to hierarchically structured, object bound functionality. Related objects in hierarchical structures are everywhere around us: for example, a kitchen contains objects related to cooking. Depending on individual sorting preferences, a kitchen cabinet may be dedicated to tableware, with plates, bowls and cups on different shelves and a drawer containing forks, knives and spoons in different segments.

And a *pointing device*, which is nothing more than the extension of the hand, seems like a *natural* way of interacting with objects such as files. Although they cannot be physically touched, through the pointing devices they can be manipulated.

It seemed that different principles were at work in the case of the coherent use of graphical user interfaces. Normally, sticking with an approach that is well researched, has evolved for over three decades, and is widely known by customers, seems like a safe bet for producers of hardware and their interface designers. When customers can apply their mental model of how a familiar

[2] Or, in the case of an optical storage device, stored by deformities on the surface of a circular disc.

system works to how a new device works, the transition from one to the other is made much easier. Also, customers are not alienated by new features but can value them in the context of the familiar as improvements and additional benefits.

While GUIs had clear advantages in desktop computing [3], lack of market penetration or *critical mass* for interfaces using different sensory modalities may have been one reason that lead to GUIs becoming a self enforcing standard whereby they are even utilised on devices where their advantages turn into disadvantages.

A second reason for the dominance of GUIs on handheld devices may be that the technologies required for alternative interfaces were not sufficiently evolved to be implemented in consumer products. On the other hand, it could be argued that *form follows function* and if producers had been keener to employ these sensor based technologies more emphasis would have been put on their development. For example, it is obvious that more emphasis has been placed on visual rather than audio interface design for mobile devices when the release cycles of the 3D video library OpenGL is compared with those for 3D audio libraries OpenAL/SL for the Android OS; While OpenGL ES 1.0 has been available on Android since version 1.0, OpenSL ES 1.0 has only been supported since version 2.3.

A third possible reason for why we do not see more alternatives to GUIs has become the motivation for this thesis: There were many inspiring ideas, concepts, and prototypes of mobile audio interfaces[4] such as Mynatt et al.'s Audio Aura [13] or the oft-cited Nomadic Radio [14] by Sawhney & Schmandt. Many researchers addressed one particular aspect of user interaction, such as Williamson et al. [15] who found a compelling way to display the number of messages received or Pirhonen et al.'s Touch Player [16] that allows a music player to be controlled by gestures on a touch screen. However, so far no holistic concept has been developed for an eyes-free interface that takes into account the tradition of metaphorical thinking about comput-

[3] The advantages and disadvantages of visual interfaces are discussed in greater detail in section 2.1.6 of chapter 2.

[4] Please refer to section 2.6.5 of chapter 2 for an extensive list of interfaces utilising sound.

ers and the subsequently derived requirements expressed through the WIMP paradigm. Such a holistic design concept for an eyes-free interface for mobile devices would need to support:

- Windows
  - a way to group objects
  - a way of content retrieval
  - a way to focus attention
- Icons
  - a way to present objects and containers as entities
- Menus
  - a way to structure objects and containers
  - a way to gain an overview of structures, items, and options
- Pointing Device
  - a way to navigate through hierarchical structures
  - a way to select and manipulate objects

The research presented in this thesis seeks to fill this gap. While focussing on one particular aspect at a time, step by step components for a holistic design concept for an eyes-free interface enabling a user to successfully interact with a mobile device are evaluated, and design recommendations are put forward. Towards the end of this thesis in chapter 8 the design concept is described as a whole. However, completeness of the factors researched and the components integrated into the design concepts is not claimed; other systemic orientations and solutions are feasible.

1.1 Research Approach

This dissertation has its focus on human-computer interaction research and practices. In an attempt to reconcile the conceptual metaphors that dominate human understanding of computer systems with intrinsically non-visual forms of human-computer interfaces, the work in this thesis focuses on the development of a holistic, non-visual interface design concept. This task is addressed by exploring the following research questions. For the interface design of mobile devices:

RQ 1 What are the advantages and disadvantages of using sound?

RQ 1.1 How can spatial sound be utilised?

RQ 1.2 What are the advantages and disadvantages of using spatial sound compared to stereophonic or monophonic sound?

RQ 1.3 How can acoustic distance perception be used as an aspect of interface design?

RQ 1.4 How can acoustic distance perception be improved?

RQ 2 What are viable non-visual multimodal interaction techniques?

RQ 2.1 What are the advantages and disadvantages of different tactile interaction techniques?

RQ 3 What is a good way to help users obtain an overview of available items and options?

RQ 4 Which interface metaphors fit the design space and comply with the WIMP paradigm?

RQ 5 How can the focus of attention be supported?

To meet the objectives and research questions stated above, a thorough literature review was completed and several user studies were conducted. Four prime data gathering methods were employed: Data logging on experimental devices, questionnaires, interviews, and observations by the researcher who conducted the study. The data were analyzed with the aid of a computerized qualitative data analysis program, interpreted, and discussed.

1.2 Research Contribution

The research contribution of this thesis is twofold: Firstly, the knowledge gained and the design guidelines derived are not only relevant in the context of the proposed design concept but are also applicable in a much wider

context. Each study's individual contribution is stated in the corresponding chapter. The individual contribution of each study conducted for this research is stated in a corresponding chapter.

Secondly, this work is novel in that it tackles the problem of combining adequate presentation and interaction methods into the holistic design concept for an eyes-free interface, summarized in chapter 8. While the concept builds on the tradition of eyes-free and auditory interface design and therefore incorporates many contributions from previous research, it is mainly developed along the lines of the theoretical reasoning presented and discussed in chapter 2 and the results gained in chapters 3 to 7.

1.3 Structure of the Thesis

Chapter 2 addresses RQ 1, 2, 3, 4, and 5. Theoretical frameworks from disciplines relevant to the work presented are introduced and the required background knowledge is reviewed and analysed. This chapter also presents introductory literature reviews, while more subject-specific reviews can be found in the respective results based chapters. Also, an overview of multimodal and eyes-free interfaces for mobile devices is given, this is followed by a discussion of previous research on non-visual techniques, proposed solutions to problematic issues, and, finally, crucial factors for future development are identified.

Chapter 3 addresses RQ 1.3 and 1.4. In the context of spatial cognition tasks two methods for the display of distance are evaluated in a user study. Their applicability in the context of an eyes-free interface is discussed and design recommendations are given. The insights gained from this study are incorporated into the design of the prototype interfaces introduced in later chapters.

Chapter 4 addresses RQ 1.1, 1.2, and 3. Obtaining an overview of available options or items contained in a folder is a core aspect of user interaction. When presenting many items, the serial nature of sound poses challenges regarding the display duration, promptness and comprehensibility that must be addressed. This is done by means of a user study

and an evaluation of the data thus generated.

Chapter 5 addresses RQ 1. Simulator or motion sickness has been proved to be a source of confusion in visual 3D interfaces. This chapter presents a study which was designed to identify the occurrence of similar effects for interfaces utilising 3D sound.

Chapter 6 addresses RQ 1.2. As mobile phones are first and foremost communication devices, the impact of different sound reproduction methods is evaluated in the context of speech based synchronous applications. The results of a user study are described and their implications for the design concept are discussed.

Chapter 7 addresses RQ 2, 2.1. and 5. In this chapter various methods of user interaction with eyes-free interfaces are proposed, evaluated, and discussed. While, in an initial study, the design space of gesture based interaction techniques is explored, a second study focuses more closely on their applicability in the context of an auditory interface on a handheld device. A third study is focussed on the exploration of the impact on interaction with eyes-free interfaces in a semi-realistic scenario where another demanding primary task is ongoing. The results of each study are discussed and design recommendations are derived.

Chapter 8 addresses RQ 4 and proposes the holistic design concept. The knowledge gained and the guidelines derived in previous chapters lead to the development of a holistic, non-visual interface design concept, which is presented in this chapter. Also, the discussion of the impact of conceptual metaphors on interface design touched on in chapter 2 is resumed and incorporated into the proposed design solution.

Chapter 9 The dissertation closes with a critical appraisal of the work presented and an outlook to future research activities.

Chapter II

Theoretical Background

The previous chapter introduced the main research focus of this thesis, namely the potential of eyes-free interfaces utilizing spatial sound in mobile human-computer interfaces. In order to design, implement and evaluate such interfaces, it is first of all important to comprehend the progress that has already been made in this field of research and development and, to identify the contributions made by psychoacoustics (section 2.1), psychology (section 2.2), and presence research (section 2.3) to this field. This chapter aims to provide an introduction to those subject areas which play a key role for the research presented in later chapters. Issues of simulator sickness (section 2.4) are also addressed and a brief overview of spatial audio in augmented reality applications is given (section 2.5).

The chapter concludes with a detailed look at components of interfaces implementing sound as the main modality, such as auditory icons and earcons (section 2.6.1), how spatiality is used as an interface dimension (section 2.6.2), and how issues of multitasking and simultaneous source presentation are addressed (section 2.6.3). A reflection on the use of metaphors in auditory interfaces is particularly emphasised (section 2.6.4). This area has been the subject of very little research and the oversight is clearly apparent when a comparison with the volume of literature available in the study of psychoacoustics is made. Finally, a broad overview of applications utilizing sound is given (section 2.6.5), these are grouped into applications that use sound for a particular aspect of the interface or interaction, for a particular application area, or for a special user group.

2.1 Psychoacoustics Overview

Psychoacoustics is the study of the subjective perception of sound. The study of psychoacoustics covers how we interpret what we hear, how we localize sound sources, how we distinguish between several sound sources and how certain factors affect our capabilities to achieve these tasks. A moderate knowledge of psychoacoustics is mandatory for any auditory interface designer in order to understand and foresee the behaviour of the interface under different conditions or in combination with other interfaces.

For humans, the sense of hearing or auditory perception is one of the most important senses necessary for everyday life. After a brief introduction to Sound (section 2.1.1), Spatial Hearing (section 2.1.2), Auditory Memory (section 2.1.3), and Distance Perception (section 2.1.5), a comparison of visual and audio capabilities as well as a discussion of their advantages and disadvantages (section 2.1.6) will complete the overview of psychoacoustics.

2.1.1 Sound

Sound is a mechanical wave caused by the vibrations of an elastic medium or object. Sound waves consist of compression and rarefaction of the molecules comprising the medium the wave propagates through. This is usually air, but it can be any solid, fluid, or gaseous material. The perception of sound begins with the arrival of sound waves at the ear drums. The eardrum passes the oscillating variations of air pressure through to the inner ear where they are converted into electrical signals. These signals are ultimately coded into a pattern of neuronal spikes that the brain is then able to interpret.

Two major parameters describing any sound wave are frequency [measured in hertz (Hz)] and amplitude or intensity [usually measured in decibels (dB)]. The human ear can normally hear sounds with frequencies from 20 Hz to 20 kHz. The upper limit of perception decreases with age and to approximately 16 kHz [17]. Sound with frequencies below 20 Hz can be perceived by the body's sense of touch. The audible threshold of sound intensity or sound pressure level is $2\text{x}10^{-5}$ Pa, which is defined as 0 dB SPL. The upper limit is not clearly defined, but any sound with a high intensity (above 100 dB) can cause pain and may also permanently damage of the eardrums [18].

2.1.2 Spatial Hearing

The spatial location of a sound as a display dimension can provide information about the location of a target, and can increase the number of sources that can be displayed. The two most important cues for determining a sound source's position in space are the interaural time difference (ITD) and the interaural level difference (ILD). The differences in arrival time and pressure at the two ears enable the brain to calculate the (approximate) position of the source [19, 20]. Additionally, the folds of the pinna (outer ear) cause small time delays 0-300 ms that alter the spectral content of the sound source [21]. The asymmetry of the pinna's shape causes this spectral modification to change in relation to the sound source position [20]. The pinnae only have a significant effect on sounds with a frequency greater than 4 kHz and they are effective for determining both the azimuth and elevation of these sound sources [22]. The direction and distance that sound has to pass through or around the head to reach the far ear creates the so called "head shadow", which affects overall intensity by about 9 dB. The head itself also acts as a linear filter, which varies with the position of the sound source [23]. Although less important than the previously mentioned cues, echoes from the shoulder and upper body reach the ear with an elevation dependent delay which provides both elevation and azimuth information for frequencies approximately 1-3 kHz [24]. Head movements can improve localization and reduce the number of front/back reversals [25, 26] which happen when a source in front of the subject is falsely localized to the rear or vice versa. Visual cues can help listeners to determine where a sound is located [27]. Listeners may even ignore auditory directional cues if they disagree with the visual cues [28]. Finally, early echo responses, i.e. the echoes heard in the first 50 ms to 100 ms after a sound starts, and reverberation, i.e. reflections from the surrounding surfaces, support determining the distance and direction of a sound [28].

Humans can perceive sound coming from any direction; however, the localization accuracy depends on the spatial origin of the sound in relation to the position of the listener. The localization accuracy is much higher in the horizontal plane than in the vertical plane: The angular resolution is about 1-5 degrees of azimuth in front of the listener and up to 20 degrees for

more peripheral and rear positions depending on the characteristics of the source and the presence of distracters [29, 18, 30].
The auditory resolution or the capability of separating between two or more sound sources in near proximity is approximately 2 degrees to 3 degrees in the horizontal plane and 6 degrees to 8 degrees in the vertical plane.

The interstimulus onset interval (ISOI), i.e. the delay between the onset of a first (lead) and a second (lag) source, plays an important role in the correct identification of sound sources and their position. If the ISOI is less than 100 ms, listeners have difficulty identifying two spatially separate sources and perceive one moving source instead [31]. When ISOI are increased to 150-300 ms the minimum audible angle (MAA), i.e. the just distinguishable horizontal angular deviation between two sound sources, is decreased from 4.7 degrees to 1 degree [32, 33].

For positions in space where sources produce equal ILDs or ITDs, listeners can usually successfully distinguish whether a source emanates from the left or right, but not whether it comes from the front, from behind, above, or below them. These positions with virtually equal interaural cues are perceived on an external conical surface extending from the listener's ear, producing what Woodworth referred to as the "cone of confusion" [34]. Moving the head will disambiguate ILDs and ITDs for sources located on the cone and hence resolve the localization ambiguities [25, 26].

A similar confusion caused by equivalent interaural cues is called the front/back confusion. In this case, a source in front of the subject is falsely localized to the rear or vice versa. The wrongly assumed positions tend to lie in mirror symmetry with respect to the interaural axis. Front/back confusions occur in 2 percent to more than 20 percent of subjects depending on the form of the stimulus [19].

2.1.3 Auditory Memory

Auditory memory can be distinguished as a pre-perceptual auditory memory (PPAM) and a post-perceptual short-term auditory memory (STAM). Although the temporal order of two spectrally remote short sounds can be reliably identified as soon as the ISOI exceeds about 20 ms, the duration of

the interval can have strong effects on the perception of both sounds. When a short sound is presented to a listener, a representation of this sound is initially stored in PPAM. The time available and needed for an optimal perceptual analysis of the sound is about 250 ms – the fixed temporal span of PPAM. If a second sound is presented to the listener in less than 200 - 250 ms after the first, the perceptual analysis of the previous sound is interrupted and the correct identification of the first sound is also hindered [35].

2.1.4 Masking

Masking effects can cause a reduction of audibility and hinder the perception of individual sound sources. Multiple sources, presented simultaneously (simultaneous masking) or with very short onset intervals (temporal masking), can interfere with one another in several ways. The amount of temporal masking is a function of the time gap between the signal and the masker: the smaller the inter-stimulus onset interval, the more masking occurs.

A further distinction is made between "energetic" and "informational" masking. Energetic masking occurs when a simultaneously presented signal and masker contain energy in the same critical bands. The larger the critical bandwidth (auditory filter within the cochlea) the lower the signal-to-noise ratio (SNR) resulting in stronger signal masking. With increasing masker intensity masking patterns become asymmetrically wider. The "upward spread of masking" occurs when the masker intensity is raised, and results in a considerable spread of the masking effect upward in frequency but only a minimal effect downward. High-frequency maskers are only effective over a relatively narrow frequency range in the vicinity of the masker frequency. In contrast, low frequencies tend to be effective maskers over a very wide range of frequencies.

Higher-level "informational masking" occurs when the signal and masker are both audible but the listener is unable to distinguish which elements of the perceived sound belong to the signal and which to the masker. Informational masking is greatest when the masker contains 20 or less pure tone components or when the listener is uncertain about the characteristics of a target source [18, 36].

2.1.5 Distance Perception

There are a number of potential cues to judge the absolute distance of a sound source. The first is the intensity of the direct sound reaching the listener. In an anechoic environment, the sound pressure of a spherical wavefront radiating from a point source will obey the inverse square law: it decreases by 50 percent (6 dB) as the distance is doubled[1] [37]. However, more than the expected decrease of 6 dB is required to perceive a doubling of distance so that apparent distance underestimates actual distance [38]. The distance of sound sources located farther than 1 m from a listener is often substantially underestimated while the opposite is the case for very close sound sources [39, 40]. Zohorik et al. [40, 41] suggest the exponentially compressive power function:

$$r' = kr^a$$

with $k = 1.0$ and $a = .32$ to be a conclusive formal description of the distance perception bias that he and other researchers measured.

If a listener is familiar with the sound source, changes in intensity are a feasible absolute distance cue. For unfamiliar sources, changes in the sound level are only helpful for determining the relative distance, i.e. compared to another sound source or a prior position [42].

In contrast to anechoic listening environments, in everyday listening situations reverberation and interfering sounds play an important role. Sounds reaching a listener in a real room consist of both the original sound from the source and its reverberation, composed of multiple reflections from surfaces or objects within the room. In a reverberant context, the change in the proportion of reflected to direct energy functions as a stronger cue for distance judgements than intensity changes[2] [29, 20, 38] and reverberant settings generally improve distance judgements compared to anechoic settings [18]. When sound travels over relatively long distances high frequencies are attenuated more strongly than low frequencies [43]. This phenomenon is due to

[1] If the sound source is not omnidirectional but instead a line source, like a moving car, an intensity reduction of 3 dB is commonly used in noise-control applications.

[2] Interestingly, Zahorik [38] found that for speech signals intensity has a greater perceptual impact than the direct-to-reverberant energy ratio.

absorption. If the sound source is less than 1 m from the head, low frequency ILD significantly contributes to distance perception [18].

Using distance as a dimension in auditory interfaces can be problematic due to a range of factors. One of these factors is how we use a sound source's loudness, i.e. the interpretation of the perceived magnitude of its intensity to judge its distance. The inverse square law predicts sound intensity reduction with increasing distance, but to accurately judge this distance the initial intensity of the source has to be known. Otherwise the listener cannot tell whether the sound they hear is low in intensity because it has travelled over a certain distance and already lost a lot of its intensity or whether it is fairly near and never was very intense. Naturally, the sound source intensity has to fall between the absolute threshold of hearing (measured at 20 micropascals or 0 dB) and between the threshold of pain (around 120 dB). Given that, for example, human speech has an intensity of around 60 dB when measured at a distance of 1 m, even under ideal listening conditions it will become inaudible when the distance is increased to 1 km (-6 dB per doubling of distance). As the source's intensity cannot be dynamically adjusted to cover a wide range of distances, using a simulation of a realistic sound field would restrict either the displayable distances or the displayable sound sources to a much greater extent than is acceptable.

Ultimately, the designer of an auditory display has no control over the intensity at the eardrum of the receiver, since the user will always have final control over the overall sound pressure level by adjusting the volume on the playback device.

2.1.6 Hearing vs Sight

There are some important differences between auditory and visual perception, which dictate and limit the use of each sense in human–computer interfaces. Human eyes perceive approximately 80 degrees in the horizontal and 60 degrees in the vertical dimension from a central point of view [44]. This is the window through which we see the world. The resolution or the focusing area decreases from the center of view to the peripheral areas. The high focus area is only approximately 2 degrees around the central area. There is

no such limitation for sound.

Another difference between the auditory and visual perceptions can be best illustrated with the comparison between the parallel and serial communication channels. The visual channel has a much higher information processing ability, i.e., 4.32×10^6 bits/sec [45], equalling approximately a 1024×1024 pixel bitmap image with 256 colors, thus enabling a great amount of information to be perceived simultaneously. The auditory channel is much narrower with a bandwidth of 9.900 bits/sec [46], which results in a slower processing of presented information auditorily. Humans are primarily visually oriented. Our sight is fully occupied most of the time and is therefore relatively insensitive to minor changes in the viewed image. Sound, on the other hand, can rouse our attention at any time and can be used for delivering temporally important information. The auditory system is particularly well-suited for alerting and monitoring, due to its ability to ignore expected sounds and rapidly detect unexpected sounds [47]. Response times to auditory stimuli are often lower than to visual stimuli [48].

Another major disparity between the two channels is the transience of information presented either visually or by sound. Visual information usually stays on the screen for a longer period, while auditory information is delivered sequentially and hence has to be replayed if not remembered or understood right away.

2.2 Attention & Distraction

Users of mobile devices often need to focus on several tasks in parallel. Questions concerning the varying degrees of attention required for a task are equally relevant here for the psychoacoustic dimension of using spatial sound in human-computer interfaces; these include the amount of distraction caused by competing tasks, the display technique, or disturbance factors. This section will give a brief introduction to the overall concept of attention and the specific implications that may apply for auditory interfaces. Attention is commonly defined as concentration of awareness to a specific source of information or a phenomenon to the relative neglect of other stimuli [49][3], whereas

[3] cf. `http://www.britannica.com/EBchecked/topic/42134/attention`

distraction is the diversion of attention from a chosen source of information onto one or several other sources. Attention can either be directed willingly or it can be evoked instinctively by a key stimulus, such as an alarming sound or a fast moving object.

According to the multiple resource theory of attention [50, 51], humans only have limited amounts of attention available at any given time. Different tasks can use different attention resources or share them. If the performed tasks rely on the same resource, they can interfere with each other and affect the performance. For example, driving a car is visually demanding. Operating a navigation system or mobile phone with a visual interface competes for the same resource associated with visual perception and can therefore cause distraction from the primary (driving) task [10, 52, 53]. Using a speech-based interface on the other hand demands resources associated with auditory perception, and would not compete for visual attention; as such the utilization of a speech-based interface would be less detrimental for visual attention [54].

Research [55, 54, 56, 57] has shown that the complexity of the competing tasks plays a key role a person's ability to pay attention to a given task. To stay with the example of driving a car, Young et al. [58] have shown that physical and cognitive distraction significantly impair the a driver's visual search patterns, reaction times, decision-making processes and the ability to maintain speed, throttle control and lateral position on the road. If the primary task is of low complexity attention can be safely diverted to a secondary cognitive task. The perceived complexity of tasks depends on, amongst other things age, emotional state and experience [59].

As an alternative to the multiple resource theory, investigators have proposed that attention is comprised of an assortment of skills [60, 61], such as the ability to segregate two messages from one another. According to this theory, difficulties with multitasking are caused by the inability to segregate different input channels and to keep "one line of parallel processing from interfering with another" [62]. Attending simultaneously to a visual and auditory stimulus is easier than attending to two visual or two auditory stimuli because selection among competing inputs depends on the ease with which they can be discriminated. According to Hirst the two consequences of channel interference are either a confusion of stimulus responses (reac-

tion intended for stimulus A follows stimulus B) or a complete breakdown of multiple task performance [62].

As Hafter et al. [63] point out, auditory attention has often been studied in terms of a listener's ability to detect signals from background noise and to focus on the expected occurrence of a signal along a monitored dimension. If the primary task is auditory in an environment with a lot of auditory distractions or noise, like listening to one conversation in a multi-source environment, listeners have the ability to filter simultaneous sounds and to concentrate on only one. This phenomenon has been coined the "Cocktail Party Effect" by Cherry [64], see also [65, 66, 67, 68], and has lead to the Filter Model of Attention proposed by Broadbent [69]. The premise of the Filter Model is that physical attributes of the input message guide the sensory register's pattern recognition and lead to the focus and continuation of attention.

In a multi-talker situation, in addition to the abilities of the listener, factors such as the distance of the speakers, the level they are speaking at, the characteristics of their voice, their gender, which way they are facing, etc. all influence how well a listener can concentrate on one or several voices. But even if a listener is not concentrating on a particular speaker, some higher-level information, such as the listener's name, might break through to awareness [70, 71]. Treisman [72, 73] corrected the physical Filter Model towards the concept of attentional attenuators, which allow gradations and selectively reduce the amount of information passing through the senses. Deutsch & Deutsch [74] argued that selection is based on pertinence, occurs after the pattern recognition stage and includes semantic factors as well. Norman [75] complements Deutsch & Deutsch's model by arguing that both the pertinence of the sensory input and its strength play an important role in the selection process.

When designing mobile human-computer interfaces, limited attention and the continuous exertion of load during human information processing are important issues that have to be taken into account [53]. As all mobile usage scenarios are inherently multitasking situations where the user has to attend to a primary task, like walking or driving a car, and a secondary task, i.e. performing a task on a mobile device, such as writing or reading a

text message, execution of the secondary task should not distract from the primary task [76]. Therefore, knowledge about the cognitive load created by a task is crucial when designing mobile interfaces.

The following section will briefly introduce the Cognitive Load Theory (CLT) and summarize a practical approach to measuring the cognitive load that is created in dual-task conditions.

Measuring Cognitive Load

The term *cognitive load* refers to the mental resources available for completing tasks at a given time, by an individual person, and under specific social and environmental conditions. The Cognitive Load Theory developed by Sweller [77] allows predicting allows for the prediction of performance when using alternative interfaces or setups, and hence lets designers and developers minimize the cognitive load created by their systems.

The CLT is based on the Multiple Resource Theory[4] of Navon [50] and Wickens [51] and the (theorized) mechanisms of the working memory. The working memory is assumed to be capable of briefly storing and manipulating information involved in the performance of complex cognitive tasks such as reasoning and comprehension. The working memory holds the most recently activated elements of long-term memory and moves them into and out of memory storage[5]. According to Baddeley & Hitch's multicomponent model of the working memory [78, 79, 80] multiple independent processors are associated with different sensual input modes. A phonological loop briefly stores auditory-verbal information while a visuo-spatial sketchpad briefly holds visual images. Phonological loops are 'the voice inside ones head' and hold inner speech for verbal comprehension and rehearsal. Without continuous repetition acoustic information fades after about 2 seconds [81]. Added to this model at a later, the episodic buffer forms a semantic and chronological tie between information from the visuo-spatial and phonological subsidiary systems. It binds information across modalities with time sequencing to form

[4] See section 2.2 for a description.

[5] cf. `http://www.britannica.com/EBchecked/topic/1431950/working-memory`

a unitary episodic representation that makes sense to us like a scene from a play or a memory of a conversation [82]. A central executive acts as supervisory system and controls the flow of information to and from its subsidiary systems, integrates new information and initiates decision-making processes.

Performance measures of cognitive load included task completion time, reaction time, error rate, memory retrieval time and correctness, rate of physical activity and speech, spoken disfluencies, multimodal integration patterns, and other indices [83]. Subjective measures of cognitive load cannot be collected in real-time and are usually measured after the experiment through tests such as the NASA TLX [84], the Subjective Workload Assessment Technique (SWAT) [85], or VACP (Visual, Auditory, Cognitive, Psychomotor) workload modelling techniques [86]. Physiological measures for assessing cognitive load include brain activity reflected in EEGs or monitoring of pupil size, which require special instrumentation and are prone to confounds [83].

Divided attention or dual-task studies, in which a test subject solves a primary task while also completing a secondary task are a common research strategy for examining cognitive load. While many studies are conducted in a laboratory environment, others have created more complex setups that reflect the usage scenario for the mobile device/interface (see [87] for a review of methods).

2.3 Presence & Social Presence

Mobile device-based communication is evolving rapidly, expanding from voice and text, to a wide range of social media and communication features offered by applications like Twitter[6] and Skype[7] or communities like Facebook[8] and MySpace[9]. The pervasiveness of these services is evidence of an increased demand for the (technically mediated) experience of being connected to a social network, sharing a (virtual) space or being otherwise in contact. The following section gives a brief introduction to the concepts of presence and

[6] `http://twitter.com`

[7] `http://skype.com`

[8] `http://facebook.com`

[9] `http://myspace.com`

social presence and then proceeds to a discussion of the role of sound in the light of this context.

There is a comprehensive and very diverse body of research on the concepts of (tele-)presence and social presence. Short et al. introduce the concept of social presence by defining it in relation to and as a subjective quality of a communication medium as well as an

> "[...] attitudinal dimension of the user, a 'mental set' towards the medium." [88]

Biocca & Nowak [89] go further and define social presence as a temporary judgement of the nature of the interaction with the other – limited or augmented by the medium and the interaction technique. Since the 1970's the definition of social presence has been broadened to include the sense of *being with others* [90, 91], the *perceptual illusion of non-mediation* [92] and the definition of Biocca & Harms where social presence is

> "[...] the sense of being with another in a mediated environment [...] the moment-to-moment awareness of co-presence of a mediated body and the sense of accessibility of the other being's psychological, emotional, and intentional states." [93]

(Tele-)presence is a related concept and most commonly refers to the physical and spatial sense of "being there", having the perception of being *physically* present in a remote environment. Some authors make the distinction between *(tele-)presence* and *virtual presence*, where the former denotes the feeling of being present at a remote location, and the latter indicates the experience of being immersed in a virtual environment [94].

The impact of 3D technologies on the perception of social presence has been comprehensively studied for visual experiences, but investigation of the impact of such advances on multimodal or audio-only experiences has not been as thorough. Lombard & Ditton [92] include a broad review of research conducted on presence prior to 1997 and include a more specific review of the effects of sound on the sense of presence. Despite rather mixed findings they summarize in their overview, they assume it is likely that spatial audio cues increases the sense of presence. This assumption has later been supported by

experiments finding a positive influence of spatial realism and externalization cues on the perception of presence [95, 96].

Social presence is an important factor in the design of communication technology as it provides a measure of how well different technologies can channel the whole spectrum of communication cues or *channels*. Ellis & Beattie [97] list five channels of human communication:

- verbal (words, sentences and phrases)
- prosodic (rhythmic aspect of speech, acoustic modulations used to express syntactic boundaries, focus and emphasis, or emotional attitudes)
- paralinguistic (volume, intonation, speed, etc.)
- kinesis (bodily movements, gestures, facial expressions, posture, gaze, and gait)
- standing (appearance, clothing, etc.).

A correlation between the coverage of communication channels and the perceived social presence, i.e. the more channels are covered the higher the potentially perceived social presence, is a common assumption among researchers [98, 99, 100, 101].

Short et al. [88] found that a face-to-face conversation has the highest perceived social presence, followed by video conferencing, multichannel audio, monoaural audio and speakerphone conversations. Audio-only or text-based media are ranked lower as they fail to convey a number of visual cues, such as facial expression, eye gaze, gestures, and proximity. Nevertheless, as the degree to which people wish to perceive social presence, the situational circumstances, and the purpose of the interaction contribute to the choice of medium; audio-only or text-based media might sometimes be the best fit.

2.3.1 Measuring Social Presence

In order to quantify the effect of different technology on social presence a wide range of measurement techniques has been developed. In Short et al. [88] the main method for measuring social presence is the semantic differential technique [102]. Participants were asked to subjectively rate the communication media on a series of bipolar scales such as:

impersonal 1—2—3—4—5—6—7 personal

Other examples of bipolarities are:

- cold — warm
- insensitive -– sensitive
- small — large
- passive — active
- closed — open

Short et al. found that media supporting a high degree of social presence are typically judged as warm, personal, sensitive, and sociable. A range of other subjective survey methods for determining social presence have been developed [103, 91]. However, purely subjective measures may not be sufficient as subjects are mostly naive to the concept of presence and may therefore be incapable of reliably answering presence related questions [104]. Insko [105] also points out that questionnaires are subject to response bias and might produce unstable and inconsistent responses depending on a participant's prior experience (see also [106]).

IJsselsteijn et al. [107] argue that a better approach to measuring presence is a combination of both subjective and objective measures, thus yielding different but complementary types of insight into the determinants and structure of the participant's responses. Biocca et al. [108] identify several possible objective measures of co-presence:

- Attentional behaviours (such as eye fixation on the other)
- Proxemic behavior (such as movement towards or away from)
- Physiological responses (such as increased arousal)

Some of these measurements were used in the study described in 6 to evaluate the impact of different sound reproduction techniques on the perceived presence and social presence. An introduction to previous work on influence of spatial sound and the perception of social presence is given in the following section.

2.3.2 Spatial Sound & (Social) Presence

Little and conflicting empirical research is available concerning the effect of spatial sound on the sense of social presence. For example, Reeves [109] found no differences concerning presence for monaural presentations and presentations for which the dimensionality of sound was enhanced via Dolby surround sound decoding. On the other hand, Christie (1973a, as cited in [88]) found that social presence was greater on self-report measures for a "multi-speaker audio system" than a single speaker system. Regarding spatial properties of sound and immersive presence in virtual environments, Hendrix & Barfield [96] showed that spatialized sound was favoured in terms of presence compared to non-spatialized sound. Yankelovich et al. [110] conducted an audio quality assessment to determine how differences in quality impact audio clarity, a remote person's experience connecting to a conference room, and social presence. From this research Yankelovich et al. conclude that high-fidelity stereo audio improves audio clarity, helps improve a user's experience in remote conferencing tasks, and enhances a sense of social presence.

On a related note, Baldis [111] analysed the effects of spatial audio on memory, comprehension and preference for desktop based conference software. She found that spatial audio improved all measures: increasing memory, focal assurance and perceived comprehension. In particular, it was concluded that 3D audio enhanced memory of the conference, because the spatial location provided an additional clue and resulted in a more efficient use of working memory. See also [112] for an interesting discourse on the historical and socio-cultural paradigms of immersive audio.

2.4 Simulator Sickness

Simulator sickness may potentially have a negative impact on the the experience of pleasantness and, as a consequence, performance and acceptance of human-computer interaction (HCI) with systems deploying spatial sound. Simulator sickness is a form of motion sickness in which users of simulators or virtual environments develop symptoms such as dizziness, fatigue, and nausea [113, 114]. Both simulator sickness and the related phenomenon of motion sickness are difficult to measure. They are polysymptomatic and

many of the symptoms are internal, nonobservable, and subjective.

One of the most popular theories for explaining simulator sickness is the sensory conflict theory of Reason & Brand [115]. They believe that motion sickness occurs if there is a conflict between visual, vestibular, and proprioceptive signals in response to a motion stimulus. This disconcordance between the different cues leads the brain to conclude that the conflict is a result of poisoning [116]. To protect physical health, the brain reacts by inducing sickness and even vomiting to clear the supposed toxin from the body.

Vection, the illusion of self-motion, has been identified by Hettinger & Riccio [117] and McCauley & Sharkey [118] as one of the potential causes of simulator sickness. Studies concerning vection often assume a link between the vection measured and the potential for the device or environment to cause sickness. Several investigations have shown a correlation between (tele-)presence and immersion and the perception of vection [107, 119].

Vection can also occur in many real life situations – usually when an observer is not moving, but is exposed to a moving visual pattern, such as when, for example, watching a moving train through the windows of a stationary train, or, a film in the front rows of a cinema.

2.5 Audio Augmented Reality

The use of audio as a display modality for Augmented Reality (AR) dates back to 1995. In 1995 Bedersen [120] proposed a first prototype of an audio augmented reality application in the form of an automated tour guide for a museum. Visitors of the museum were equipped with a modified Sony MiniDisc player, a microprocessor, and a custom infrared receiver to determine their position within the museum exhibition. Infrared transmitters were places above each exhibit. The tour guide supported passive interaction: by walking up to a piece, visitors could trigger the playback of pre-recorded descriptions; by walking away the playback of the description could be stopped. Bedersen's method of tracking is similar to that which had been used in Audio Aura, built by Mynatt et al. in 1998 [13] at PARC. Employees of PARC would wear infrared transmitters, so-called active badges, which were de-

tected by a network of infrared receivers placed throughout the building. The positional information of each user was tracked by a location server. Changes in the location database prompted the system to send audio cues to the users' wireless headphones. These cues were sound effects, music, and voice, and would notify the user of the status of their email or the start of a meeting, or remind to complete tasks such as retrieving a book from the library. The longer a user lingered in certain area, the richer the information they received would become. Audio cues could also be triggered by artefacts. A bookshelf might, for example, provide a service user with information about recent acquisitions. The goal of this work was to "create an aura of auditory information that mimics existing background, auditory awareness cues" [13].

The Hear&There outdoor augmented reality system [121] designed by Rozier et al. in 2000 enabled users not only to retrieve location based "audio imprints" like music, sound effects, or recorded voice but also allowed for the creation of audio content. The system utilized a pair of headphones with a digital compass attached, a laptop, a PalmPilot, a high precision Global Positioning System (GPS) receiver, a battery, and a microphone. Using the information from the GPS and the digital compass, the Hear&There system was aware of the users position and their viewing direction. Unlike previous systems, in Hear&There sounds are spatialized, i.e. sound sources could be placed in three-dimensional space and were heard by the user to be coming from that particular position. In a similar approach to that utilized by Bedersen and Mynatt et al. playback was triggered by entering a zone containing one of several layers of audio imprints. See also Lyons et al. [122] for a virtual reality game in which players move around in the real world and trigger actions in the virtual game world. Lyons et al. used wearable sensors and employed RF-based locationing.

With some similarities to the Hear&There system, the 3DAAR (3D Audio AR) system introduced by Sundareswaran et al. in 2003 [123] is designed for outdoor use and supports spatialized sound. It is intended for mobile security applications, providing audible alerts to mobile users indicating the location of threats and waypoints for navigation to target locations.

The 3DAAR system utilizes GPS–based position measurement, magneto-

meter-based head orientation tracking and speech recognition.

The Mara (Mobile Augmented Reality Audio) framework proposed by Härmä et al. [124] in 2004 primarily focuses on the development and evaluation of the augmented reality audio (ARA) headset - stereo earphones with integrated binaural microphones. Through this headset, signals from the wearer's surroundings are either directly routed to the earphones exposing the wearer to a pseudoacoustic representation of the real environment or virtual sound events are mixed with microphone signals to produce a hybrid; an augmented reality audio representation. The MARA framework can be used for a variety of audio augmented reality applications. Listening tests in which subjects were presented real sounds from loudspeakers and virtual sounds through the ARA system have shown that even experienced listeners often failed to discriminate virtual sounds from test sounds coming from the loudspeakers.

The LISTEN project by Terrenghi & Zimmermann [125] follows up on [120] as another audio augmented environment for museums. LISTEN allows users to move freely in the physical space and listen to spatialized audio sequences emitted by virtual sound sources positioned throughout the exhibition. Sensors are placed in the environment and track the users movements. As in many of the systems mentioned above, the playback of specific sound information about a connected visual object is initiated when an "object zone" is entered.

Another system using approaches similar to those developed by Lyons et al. [122] is the "Roaring Navigator" by Stahl [126]. This system is based on spatialized audio navigation within a real-world zoo environment. Spatialized recordings of animal voices were used to show the location of animals and the system proactively presented detailed information about each animal.

CORONA by Heller et al. [127] is an interactive audio experience of a medieval coronation feast in the Coronation Hall in Aachen, Germany. Within the building the user can walk through ten areas each containing distinct audio events, i.e. a single character telling a story or several characters talking among themselves. The system is implemented on an Apple iPhone utilizing the inbuilt OpenAL 3D sound library. The Ubisense real time location system is used for tracking inside the building. The users' head orientation is

measured with a compass sensor mounted on the headphones.

In 2008 Woices[10] becomes the first commercial mobile phone application primarily relying on audio augmented reality. Like Hear&There, Woices supports geotagging of locations and the creation of user generated audio content to create audio guides or commentaries for specific locations.

Toozla[11], another commercial mobile phone application, employs Wikipedia[12] articles, which are converted to audio using text to speech technology. As in Hear&There and Woices, users can record their own reviews of attractions. Similar to Audio Aura, LISTEN, the "Roaring Navigator" and CORONA, in Toozla playback of audio in Toozla is triggered when the user comes with in range of a location for which content is available. The system will first offer an overview and will then provide more detail if the user stays in the area.

2.6 Auditory Interfaces

In this section, an overview of the diverse elements that comprise auditory interfaces is given. This includes object or structure representations (2.6.1), the use of spatiality as a display dimension (2.6.2), parallel source or stream presentation (2.6.3), and the use of interface metaphors (2.6.4). The section ends with an in-depth review of previous research on mobile auditory interfaces focussing on those using sound for a particular aspect of the interface or interaction, for a particular application area, or for a special user group (2.6.5).

2.6.1 Auditory Icons & Earcons

In the 1980's Gaver [128] developed the concept of using natural everyday sounds to represent events and objects in a computer interface. The so-called "auditory icons" were included for the first time in the Apple SonicFinder [129] and are still in use today. Auditory cues are usually divided into three categories based on their abstraction level: they can be iconic,

[10] `http://woices.com/`

[11] `http://toozla.com/`

[12] `http://wikipedia.org/`

metaphorical (indexical), or symbolic. While iconic representations try to acoustically reproduce an event as realistically as possible, the metaphorical or indexical auditory cues establish an analogy between an event and an associated sound. The so-called earcons have the highest abstraction level; they do not allow any semantic relation between an event and a sound, but rather assign an arbitrary audio signal to represent an event. Earcons can be designed not only to represent a single item, but also its position in a hierarchical structure [130], either in audio-only interfaces, such as telephone-based interfaces [131, 132, 133] or in multi-modal interfaces [134, 135]. It has been shown that earcons can successfully improve the usability of multi-modal interfaces for mobile use [16].

Almost any listener can easily interpret simple auditory icons representing an event or object by playing a typical sound (e.g. deleting or "throwing away" a file being represented by the rattling sound of a trash bin). Developing auditory icons with a high compatibility for more abstract events (e.g. changing the active profile of a mobile phone) can thus be difficult and can, in addition, lead to a reduced ability to interpret the auditory icon without training. The meaning of earcons needs to be learned a priori and is not transferable to other earcon "languages". Comparative studies of simple auditory icons and earcons show no significant difference in efficiency between the two [136, 137, 138], including when used in combination with spoken menu items for locating different items in a hierarchical structure [139]. As the abstract auditory cues were named "earcons", the study cited above introduces "spearcons". Spearcons are audio cues generated by converting the text of a menu item to speech and then speeding up the resulting audio clip until it is no longer comprehensible as speech. Spearcons in combination with a spoken menu text show a slight advantage over the spoken only menus but a strong advantage over earcons [139].

The so-called "hearcons" were created to support the navigation of web pages [140] or hierarchical menus [141]. Hearcons are three-dimensional abstract auditory objects positioned in an auditory interaction realm. They constantly emit sound and can be manipulated with the use of a pointing device.

As has already been discussed, sound events can be successfully used to

represent events within a hierarchical structure. It is possible not only to code information about the meaning of an event but also about its position in a hierarchy. The amount of information that can be represented by the sound event itself is clearly limited, but the way of representing the event can be used to add additional information. One method of representation is the use of spatial sound. The next section will give a short introduction to spatial auditory interfaces, their potential and limitations.

2.6.2 Spatiality

Many auditory interfaces (see section 2.6.5 for an overview) use three-dimensional sound as an additional display dimension. 3D sound synthesis can effectively mimic a realistic hearing experience while listening to stereo- or monophonic sound often leads to the sound being perceived as intracranial – located inside the head.

Individual sounds or whole scenes can be effectively generated through head-related transfer functions (HRTFs) [20, 142, 143]. HRTFs are frequency responses of an acoustic path from the sound source to the human eardrums. As they take into account many of the cues humans use to localize sounds, such as reflections of the shoulders, head, and pinna, they model the human acoustic system and allow, if the listener is wearing headphones, an externalization of the sound source. HRTFs are usually measured as head-related impulse responses (HRIRs) for each individual listener separately. When generalized HRTFs – measured with a dummy head – are used for creation of virtual sound sources, higher localization error rates or the so-called *localization blur* can occur [144, 29]. Spatial sound can also be delivered through multiple speaker setups, in which an array of physical speakers is arranged around a listener. Common techniques for this are Ambisonics [145] or different panning techniques like vector based amplitude panning (VBAP) [146, 147, 148] or distance based amplitude panning (DBAP) [149][13].

[13] See [150] for a comparison of these three methods.

2.6.3 Simultaneity

In graphical user interfaces (GUIs) support for multiple item presentation and multitasking is are important features. The serial nature of sound makes these features much more difficult to design in auditory interfaces. The visual sense can process the information presented in parallel and then focus on one particular bit of information while parallel processing of auditorily presented information is more limited due to a lower resolution[14] of the sense itself and limitations arising from masking effects (see section 2.1.4). There is, however, the "Cocktail Party Effect", i.e. the human ability to selectively focus on one out of several sound streams (see section 2.2 for further information). A number of researchers have addressed the subject of simultaneous information presentation in auditory displays. Their findings and recommendations are presented in this section.

In their comprehensive work on the perception of multiple earcon display, McGookin & Brewster [151] found that varying the number of concurrently presented earcons significantly affects participants' ability to identify them. In their study transformational earcons were used, i.e. in this case earcons in which three auditory parameters, such as pitch or rhythm, were directly mapped to a data parameter. While the proportion of correctly identified earcons was reduced to 30 percent as the number of concurrently presented earcons increased to four, correct identification of individual earcon attributes was much higher, dropping to only 70 percent. Furthermore, the authors found that both modifying each earcon's timbre and playing them with a 300 ms interstimulus onset interval significantly improved their identification. In the discussion McGookin & Brewster speculate that reducing the complexity of earcons may lead to an increase in identification performance.

Studies of auditory display techniques using speech such as the work of Brungart et al. [152] have come to similar conclusions: increasing the number of concurrent audio items leads to a decrease in identification rates. Brungart et al. found that listeners' correct identifications of a target phrase decreased by approximately 40 percent for each talker added to monophonic playback of same-sex talkers over headphones. While listening to only one talker led

[14] See section 2.1.6 for a comparison of both senses.

to a performance of nearly 100 percent, performance for four competing talkers dropped to 24 percent of phrases being correctly identified. Brungart & Simpson [153] found that spatially separating the concurrent talkers can produce improvements in intelligibility. Spatial separation in distance improved performance with a same-sex speech masker by 28 percent to 33 percent depending on the normalisation technique used.

In his work on the TableVis application giving non-visual overviews of tabular numerical information Kildal [154] investigated the effectiveness of High Density Sonification (HDS), i.e., the simultaneous sonification of all or some data points in a set. He found that exploring tabular numerical data using HDS is more effective, more efficient and produces lower subjective workload than exploring it with speech. While non-speech sound lead to about 80 percent accuracy, the same information rendered in speech yielded only about 60 percent accuracy.

2.6.4 Interface Metaphors

Visual interfaces are fundamentally different from auditory interfaces. Thus, when designing auditory interfaces and interaction techniques metaphors cannot simply be transferred from visual interfaces but require careful thought and adaptation. This section will first give a thorough introduction to the origins of metaphors in literature and cognitive linguistics and will then proceed to analyse and discuss the use of metaphors in existing mobile auditory interfaces.

In linguistics metaphors are, along with metonymy and synecdoche, a subcategory of tropes - rhetorical figures of speech. A synecdoche is a play on words where a part represents the whole, as in the expression “hired hands” for workmen, or the name of the material for the thing itself (“steel” for sword). Synecdoche is closely related to metonymy, which works on the basis of the proximity or correspondence between two concepts (“crown” for queen), whereas metaphor works by the similarity between them. In a metaphor a word or phrase that ordinarily designates one thing is used to explain another, thus making an implicit comparison.

William Shakespeare is well know for his use of metaphor:

All the world's a stage,
And all the men and women merely players;
They have their exits and their entrances; (As You Like It, 2/7[15])

In this metaphoric example of the beginning of Jaques's monologue in *As you like it*, "the world" is compared to a stage and life to a play. The concept, idea, or thing that is to be explained is called the target or tenor, in our example "the world". The term, concept, or source used for the comparison is known as the metaphor's vehicle, in this case "a stage". By comparing the world to a stage, Shakespeare draws on the images evoked in the reader/listener to transport Jaques's view on life.

Every metaphor has its implications and limitations, often referred to as the metaphoric entailments. To remain with the same example, the metaphor implies that there are roles predefined by a playwright, a God perhaps, who writes us into being, that the dramatic course of the play is already set and actors are solely performing without influencing the course or outcome of the play, that actors enter and leave (are born and die) but the characters or roles are immortal and will be filled by other actors. It implies that there is an audience watching the play, which in the Elizabethan times could have been a reference to all beings above humans in the hierarchically structured system of order, i.e. angels and God. These are the entailments, the conceptual similarities, the overall metaphor is based upon; it is the ground of the metaphor.

Besides understanding metaphors only in terms of their appearance in literature or poetic language, a much wider ontological impact has been ascribed to metaphors by cognitive linguistics in the early 1980's. Lakoff & Johnson's [12] generalized definition of metaphor is that of "understanding and experiencing one thing in terms of another". This expands the scope of metaphors beyond their application as a figure of speech towards shaping the mental models of the reality we live in. Lakoff & Johnson have found that our conceptual models of the world are fundamentally metaphoric in nature. Metaphors structure what we perceive, how we interact, and how we relate

[15] `http://en.wikisource.org/wiki/As_You_Like_It/Act_II#SCENE_VII._Another_part_of_the_Forest.`

to other people. Thus metaphors are "pervasive in everyday life, not just in language but in thought and action". Two of their examples are that of *argument is war* and *time is money*. Thinking of arguing in terms of fighting is expressed in the phrases we use. For example:

> He attacked every weak point in my arguments.
> Her criticisms were right on target.
> I've never won an argument with him.

In a similar fashion, we refer to time by using words that we would use when referring to money:

> Writing that email cost me an hour.
> I've invested a lot of time in her.
> She's living on borrowed time.

In both cases a conceptual structure from another domain is applied to the domain in focus. In our two examples argument and time are the tenors or target domains, whereas war and money are the vehicles or source domains. As using the terms target and source helps to characterize the directional nature of the mapping we will use them instead of the vehicle-tenor terminology in the course of this article. If we imagine seeing argument as a dance it would change the way we think about arguments, the way we experience them, and how we talk about them. Trying to word phrases that use dance as a source domain it becomes obvious, that the way we are used to thinking about arguments is so deeply routed in our culture, that phrases like

> He supported my perspective.
> Her examples were elegant and full of verve.
> With ease we skipped through the pros and cons.

would simply not refer to arguing for us, but to something else. Often the source domain is concrete, sensate, and taken from everyday experience, whereas the target domain is rather abstract or specialized. We may have trouble grasping the meaning of time, but using money as a source domain well known to us and the ones we are addressing enables us to find imagery

and structures we can apply to something as abstract as time. Thus we can think and talk about time or similarly abstract phenomena but the concepts we draw from, like money or war, will eventually become a firmly established conceptual mapping between these domains and only be recognized as such on a closer look. These conventional metaphors are so deeply rooted in our cultural heritage that people would neither call them such nor attach any special attention to them.

Lakoff & Johnson propose a taxonomy of metaphor types. They distinguish ontological, orientational, and structural metaphors, i.e. systematic multiple attribute metaphors. Based on Lakoff & Johnson's taxonomy, Barr at al. [155] further sub-categorize structural metaphors into process and element metaphors. In the following paragraphs, these individual metaphor types are described concisely. Examples of such metaphor types and a reflection upon metaphors used in prior research conclude this section.

Image Schemas

Image schemas are commonly defined as recurring structures of experience. They shape our patterns of how we understand the world around us as well as the way we think and reason. Image schemas are acquired by experience, physical interaction with the world, and they are influenced by our cultural and historical context.

Containment, path, source-path-goal, blockage, center-periphery, link, scale, contact, full-empty, near-far, to name just a few, are image schemas (see [156] for a more comprehensive list of image schemas). Many of these schemas form the basis of metaphors used in interface design. By mapping the image schema onto a target, the experiential coherence and value judgements of the image schema is generally preserved in the target domain. This is called the "Invariance Hypothesis" [157, 158]. The process of finding a file in a hierarchical menu structure, for example, is following the experiential logic of the *source-path-goal* schema. The *source* is the current location in the structure. The *goal* is the location of the desired file. Only by following the correct *path* can the file can be obtained. As a result following the correct path and navigating towards the goal is positively valued, while diversions

or the inability to reach the goal are negatively valued "errors" [159]. More examples will be given in the following passages.

Orientational Metaphors

Orientational metaphors use the orientational image schemas of physical space. They mostly originate in how we, as bodily beings, perceive the physical world. They refer to spatial orientation such as central-peripheral, in-out, front-back, up-down, on-off, etc. Lakoff & Johnson [12] stress that how we classify and rate these physical experiences is not arbitrary but also strongly influenced by our cultural affiliation. For example, we think of increasing the volume as turning it up, a progress bar runs from left to right, objects in the centre are more important than those in the periphery.

Orientational metaphors, such as those using the up-down dualism, are highly systematic not only internally but also externally among various spatialization metaphors. This means that the *more is up* metaphor will likely be consistent with the way we apply it – we increase the volume, pile up files, have a high contrast, the battery is low, etc. But there is also an overall external coherence among ontological metaphors: As the *up* orientation refers to *more*, which generally refers to *good* and *well-being, happy is up, health is up, alive is up, control is up, and status is up* [12].

Ontological Metaphors

According to Lakoff & Johnson [12] three types of ontological metaphors can be distinguished - the entity metaphor, the container metaphor, and the substance metaphor. The use of an ontological metaphor enables us to refer to abstract concepts, quantify them or identify aspects and causes. Thinking of data in terms of an object that can be moved, copied, or named is a simple example of an entity metaphor. We use it to have a way of relating to the binary representation of information stored in patterns of magnetization on a magnetically coated surface (like on a hard disk drive) or in deformities on the surface of a circular disc (like on a CD or DVD). According to Lakoff & Johnson [12] personifications, such as the speaking paper clip representing the Microsoft Office Help System, are subsumed under entity metaphors.

Container metaphors represent certain concepts as having an inside and an outside and are characterized as being able to hold something else. A classic example for this is the way we think of folders holding files or the trash bin containing objects to be deleted.

The substance metaphor is a metaphor in which an abstraction is represented as a material. In computing, the term dataflow reveals how we think about data in terms of a liquid that can be piped or filtered.

Ontological metaphors in their universality and simplicity can often be found forming the building blocks of more complex metaphors such as the desktop metaphor. More complex attribute mappings are frequently found in structural metaphors.

Structural Metaphors

A number of structural metaphors relate to the source domain of physical space in general and build directly on its image-schematic structure. As Barr et al. [155] point out, a structural metaphor is often a concretised ontological metaphor. Whereas ontological metaphors are based on structurally uncomplicated and simple physical notions (*data is an object*), structural metaphors stand for more complex domains (*data is a file* or *the interface is a desktop*). To stay with the example, thinking of data not only in terms of an object but a file, more characteristics are transferred from the source to the target domain: files can be created, indexed, edited, copied, moved to the paper bin and so forth. It also allows some aspects of the data file that are more relevant to a user to be highlighted and less relevant aspects to be hidden. Although they are more complex, structural metaphors relate more directly to our experience of everyday life. We usually have a theoretical notion of what containers, objects, and substances are, but we experience and think of them in their actual implementation as apples, coins, mugs, boxes, water, or gold.

Image Metaphors

Image metaphors were later added as a fourth type of metaphor by Lakoff & Turner [160] to supplement and complete the classification. Image metaphors differ from orientational, ontological, and structural metaphors in that they map the structure, not the concept, of one domain onto the structure of another domain. This mapping can either include the attribute structure or the part-whole structure. Attribute structures are, for example, colour, physical shape, or curvature, whereas part-whole relations are, for instance, a wheel to the whole of a car or a keyboard to a computer. Thinking of a figure in terms of an hourglass or referring to a face in terms of a tomato are implementations of image metaphors. In addition, there are also image-schema metaphors, which map limited skeletal information from the source onto the target. An example is the In-Out schema, which is used to tune in or tune out of something, pass out, space out and so forth. While structural metaphors map a rich knowledge structure from a source onto a target, image-schemas represent the way people perceive the world. As described above, such images are merely general elements, but which often form the basis of more elaborate metaphors.

Process and Element Metaphors

Whereas Lakoff & Johnson define metaphors from a cognitive linguistics' perspective, Barr et al. adopt and expand Lakoff & Johnson's taxonomy to fit for the analysis of human-computer interfaces [161]. They mostly base their study on user-interface metaphors relating to Lakoff & Johnsons work, but extend their proposed taxonomy by adding process and element metaphors as subcategories of structural metaphors. Process metaphors explain how aspects of system functionality work by comparing it to processes the user is familiar with.

Many online shops resemble a real world shopping process - first items are put in a shopping cart, then the customer proceeds to the checkout, and finally pays. The icon of the shopping cart is what Barr et al. call an element metaphor. It is used to cue the user into which process metaphors are available. Element metaphors can be graphics, sounds, text, touch, etc.

Examples of Metaphors Used in Auditory Interfaces

In human–computer interfaces, metaphors are widely used to represent abstract elements of the computer system with the processes, objects, or concepts from a domain that the user is more familiar with. A metaphor can either set the whole "theme" for an interface and hence influence all of its elements, or it can apply only to a certain element like, for example, the interaction mechanism.

One of the best-known interface metaphors is the *desktop* metaphor first introduced by Alan Kay at Xerox PARC in 1970. It is a classic example of a structural metaphor, in which *files* lie on the *desktop*, are sorted in *folders*, and can be printed by dragging the files' icons onto the *printer* icon, or they can be deleted by dragging the icons onto the *paper bin* icon and so forth. Contrary to the well-established metaphors for the visual interfaces, auditory interfaces are at this time still lacking a dominant interface metaphor. In the following I will give a short overview of the variety of metaphors used in audio interfaces.

An early and influential application was Audio Windows created by Cohen & Ludwig [162]. Audio Windows is making use of several ontological metaphors: the *System is a room/container* and *sound cues are objects*. Audio Windows integrates spatial sound, audio highlighting, and gestural input recognition. Users can manipulate items by using a data glove to point at or grab items in a 3D sound space. Cohen and Ludwig applied different filters, like thickening and self-animation (chorusing or pitch-shifting, frequency-dependent phase distortion) for highlighted items.

Environmental Audio Reminders (EAR) is a system developed by Gaver [163]. The structural auditory metaphor used in EAR was the *system is an office*. For instance, the arrival of a new e-mail is signalled by a sound of a stack of papers falling to the floor - an auditory element metaphor. The stereotypical audio representations were perceived as especially intelligible. They were also easily accepted, as they did not require memorizing the meaning of a number of abstract sound cues. See also [164] for using a multiple room metaphor.

Audio Aura is an information system developed by Mynatt et al. [13]. Al-

though it is similar to EAR, is has been designed with a mobile user in mind. Audio Aura users wear wireless headphones and their movements within a building are being constantly tracked by infrared sensors. If a user enters another office or the common room, auditory icons notify them about events that occurred during their absence, such as incoming email or upcoming appointments. In addition to spoken cues, auditory icons are designed to represent natural scenes. One of the structural metaphors used is the *system is a beach*. In this case, emails are represented by seagull cries, senders by beach animals (like seals or certain birds), group work as wave sounds, and so on. It is debatable whether the auditory icons, especially the beach-themed sounds, used by Mynatt et al. are actual metaphors or rather arbitrary mappings. The lack of tenor-vehicle associations - there is no conceptual connection between emails and seagull cries - may explain why users had some problems remembering the meaning of the auditory icons. Nevertheless, the general reaction to Audio Aura was positive and users perceived it as unobtrusive.

Several researchers have used the the *system is a ring/dial* metaphor for designing their auditory interfaces. In addition to the interfaces described in the following paragraphs see also [165, 166, 167] for work implementing the ring metaphor.

Kobayashi & Schmand [168] built an egocentric dynamic soundscape to create a browsing environment for audio recordings. This application makes use of orientational and structural metaphors. A *speaker*, representing a sound stream, orbits the user's head and hence maps advancing within the audio source to movements on the circular path. Using a touchpad the user can interact with the system to either create a new speaker or switch to another already created speaker. There can be up to four speakers simultaneously playing different portions of the same sound stream.

Sawhney & Schmand [14] created the nomadic radio, a spatial audio framework. The system notifies the user of current events such as incoming e-mails or voicemail, current messages and calendar entries. Confirmation, cancellations and status are also represented by sounds. The structural metaphor of a *clock* is used by positioning audio messages in a circle around the listener's head according to their time of arrival. The user interacts with the nomadic radio through voice commands and tactile input. In both

Kobayashi & Schmand [168] and Sawhney & Schmand [14] the ring metaphor is used in the sense of *the ring is a time measuring device/clock* as they both map areas on the ring to fixed, respectively relative, times.

Frauenberger & Stockman [169] used the ring as an ontological container metaphor. The user is positioned in the middle of a virtual room with a large horizontal dial in front of them. Menu items are synthesized speech and presented on the edge of the dial facing the user while most of the dial disappears behind a *wall*. As the user can turn the dial in either direction by using a gamepad controller, the metaphor used could be the *dial is a shelf or map stand*. Only the items in front of the user can be selected or activated.

Crispien et al. [170] developed a generic spatial auditory environment for navigating between, and selecting from up to twelve sound sources. The interface is designed for aligning both non-speech and speech audio cues in a ring circling around the user's head. Only three out of twelve objects are played at the same time to create an auditory focus area. *Objects* can be reviewed and selected by using 3d-pointing hand gestures, speech recognition input, and head tracking. The focus area can be changed by "looking" at a different section of the ring or by turning the ring through commands. Objects no longer part of the focus area smoothly fade-out and adjoining objects fade-in. The generic nature of the development makes it difficult to specify the *dial* metaphor. For the selection mechanism, though, the metaphor could be that of a *directional microphone* or *spotlight*. See also Schmandt [171] for a successful application of the *fisheye* metaphor.

Pirhonen et al. [16] explored gesture and non-speech audio as ways to interact with a mobile music player. They made heavy use of orientational metaphors mapping left – right (back – forward) and up-down (more – less) movements to the functionality of an audio player. Their evaluation showed that a gesture and audio based version significantly improved the usability compared to a pen and visual based version of the same player. Pirhonen et al. demonstrated, that careful selection and evaluation of an interface metaphor can be a crucial factor for user performance and satisfaction with an interface.

Shoogle [15] demonstrates one of the most consistent uses of ontological metaphors. The mobile phone or PDA becomes a container, a *box*, that

can be shaken to reveal its contents. Elements, like messages, become *balls* that bounce around inside the container. Through acoustic and vibrotactile feedback, the authentic behaviour of objects moving around inside a container is simulated and, from everyday experience, the user can deduce the relative amount of items.

Another example of the success of a well chosen and consistent interface metaphor is SensorTune [172], a mobile interface designed to help non-expert users set up a wireless sensor network. SensorTunes's designers applied a process metaphor, that of *tuning an analog radio*: if the signal is perfectly tuned, the audio is clear, when the signal is weak, the audio output is distorted. A comparative study showed shorter task completion times for the auditory interface compared to a GUI. Also, most participants of the study found the audio interface easier and more efficient to use.

In some applications the use of metaphors is inconsistent. Often ontological metaphors are used for one part of the interface that does not match the structural metaphor used for another element of the interface. If sound streams are *objects* positioned on a *dial or ring*, then what kind of dial/ring is it? A clock? A postcard display rack? A sushi circle? Or an asteroid belt? What are the objects on that dial and how can they be manipulated? Using a *room* metaphor, for example, may help to gain a basic understanding of the architecture of the interface, but specifying what kind of room and which features it has, a *station concourse* or a *classroom* or a *lounge*, would help the user to understand what the application can be used for.

Metaphors that suggest head-, marker-, or complex gesture tracking may not be suitable for mobile interfaces, as they may potentially distract the user or may involve behaviour that is considered inappropriate in public places. The same applies to metaphors with a sound design that mirrors the environment it is used in - a beach metaphor may be appropriate for users in an office, but not for those working on a ship.

2.6.5 Mobile Auditory Interfaces - An Overview

As is the case with desktop and notebook computers, most interfaces for mobile devices are multimodal with a strong emphasis on visual represen-

tation of data. In most commercially available user interfaces for handheld devices, sound plays a minor role and is most commonly used in the form of auditory icons, either to notify and warn or to give feedback about the user's input. Although auditory interfaces have been the subject of research for more than 20 years, only a few techniques have been integrated into commercially available products. Some of these techniques and systems, a few of which have been mentioned earlier, will be described in detail in the following sections. The following catalogues organized examples according to their scope of application into techniques focusing on:

- A particular aspect of the interface or user interaction (e.g., navigation, selection, object manipulation)
- A special application (e.g., calendar, multiparty conversation)
- Blind users as a special user group

Particular Aspect of the Interface or User Interaction

VoiceNotes VoiceNotes [173] allows the creation, management, and retrieval of voice snippets or voice notes. Voice notes are recorded and organized into categories. They can contain thoughts, ideas, reminders, or lists of any sort. A set of simple commands grants access to categories and notes listed within. The system uses voice recognition to process user input. The user can also interact with the system by pressing buttons on the device. Speech output is used to give feedback or to read back the content of a voice note. Besides auditory icons, such as a page-flipping sound which indicates movement between the notes, the application also uses beeps, e.g., before and after recording, to speed up the interaction and make it less intrusive. Both microphone and loudspeakers are integrated into the prototype's hardware.

Audio Hallway Audio Hallway [171] is a method for browsing the collections of sound files. It makes use of the hallway metaphor, with rooms on both sides representing collections. A spatialized auditory collage of what Schmandt calls "braided audio" emanates from each room. While "walking down the hallway" and passing the rooms, the user hears an

acoustic indication of which collection the room represents. All sounds are rendered spatially. The user navigates either by head movements, i.e., tilting their head forward, backward, left, or right. Upon entering the room, the user can browse the sound files arranged spatially in the shape of an arc. Inspired by the fisheye lens technique, the sound file in the user's focus is played proportionally louder to aid the selection process.

Diary in the Sky The Diary in the Sky [174] explores the technique of using spatial sound to position sound items according to their semantic content. As in the case of their prototype, the sound items represented calendar entries. The entries' time stamps are spatialized and positioned on a clock-like layout, with twelve o'clock immediately in front of the user.

Spatialized Audio Progress Bar The spatialized audio progress bar [175] indicates progress, e.g., of copying a file, with sound only. It makes use of two spatialized nonspeech sounds to communicate progress, rate, and completion. The first sound on a circular layout is located at a fixed position in front of the user and provides the reference. The spatial position of the second sound communicates both the transfer progress (through its angular position) and the transfer rate (through its speed).

Touch Player The Touch Player [16] is a unique version of a music player controlled through user gesture on a touch screen. The available functions are the basic control functions such as play/stop, next/previous track, and volume up/down. The users make sweeping gestures with their finger across the screen. The sweeping across the screen from left to right - or vice versa - skips to the next track, a single tap starts or stops the playback, and so forth. The feedback signalling successful or unsuccessful gestures is given by earcons spatialized on a horizontal line in front of the user.

Multimodal "Eyes-Free" Interaction Techniques: A circular, user- centered spatial interface is created with sound sources, mostly earcons or

speech, for the menu items representing sound files or streams [167]. The interaction techniques are head gestures or gestures performed with the mobile device. The sound cues are selected by a nodding, while other head gestures are used to move the sound cues to the front when focusing on them or move them to the rear to monitor them in the "background".

Shoogle: Shoogle transfers the haptic and acoustic features of an actual container filled with objects onto the auditory and haptic features of a PDA or a mobile phone containing messages [15]. By shaking the mobile device, Shoogle sonifies the "objects" inside (i.e., the messages), which, resembling real balls, bounce against the "virtual walls" of the device. When these objects impact, the system creates an adequate sound and the device vibrates. The impact sounds can be changed dynamically. The solution was further upgraded by assigning different impact materials to different groups of sender (family, work, etc.).

EarPod: EarPod partitions the sections of a circular iPod touch pad into menu items [176]. The touch pad is divided into an inner disk and an outer dial evenly divided into sectors. Touching a section causes the system to read out the corresponding menu item, while dragging the finger between the sections allows "scrolling" through all menu items. The users select a menu item by lifting the finger. The system uses auditory icons for user feedback. All audio output is left–right spatialized to support the mappings of the sound to the pie menu.

BlindSight: BlindSight provides a practical solution for interacting with a mobile phone while maintaining a conversation [177]. By pressing the keys on the mobile phone, the user can navigate the phone's menu and access, add, or modify calendar entries. The user's input is signalled by non-spatialized speech and earcons.

GPS based Navigation: In the same way, Kan et al. [178] obtained spatial information from a sound source, Holland et al. [179], Mariette [180], and McGookin et al. [181] used global positioning system (GPS) data to guide the user toward a certain position. In their systems, a sound appears to be emanating from a predetermined location in space, which enables the user to navigate from their current position toward their target position. The user's position in relation to the sound source is continuously updated. Mariette [180] also supported head rotations to further improve the sound source localization. Holland et al. and McGookin et al. [181] both used a Geiger counter metaphor to convey distance, with a sound repeating at an increasing rate as the user comes closer to the target.

Special Applications

Audio Aura: Audio Aura [13] was one of the first and most influential mobile auditory interfaces. The system only worked within a building that was equipped with a network of infrared sensors. The users wore small electronic tags, so-called active badges, which allowed the system to identify and track each user individually. When the user entered a certain area, individualized audio cues were triggered and sent to the user's wireless headphones. These auditory cues were, for example, a summary of the newly arrived e-mails or a reminder for a meeting that is about to start. The auditory cues could also be triggered by various objects, such as a bookshelf, and contain a message about, for example, recent acquisitions. The auditory cues were not spatialized. They were designed to stay in the auditory periphery in order not to sound alarming or to draw too much of the user's attention. To achieve this effect, different sound environments or ecologies were created, e.g., the beach, in which particular sets of functionalities were assigned to various (in this case, beach) sounds. For instance, the number of seagull cries signalled the number of new e-mails, while a group activity was represented by waves – the more activity, the louder the waves. Thus, Audio Aura was the first personalized mobile auditory information sys-

tem. The system supported only indirect physical interaction, since the users had to physically enter a certain area to activate an audio cue. However, the system also changed its state in accordance with certain events such as the amount of new e-mails.

Nomadic Radio: Also one of the first, yet one of the most sophisticated interfaces is the Nomadic Radio [14]. The Nomadic Radio consisted of a shoulder-worn speaker and microphone unit and the infrastructure supporting content retrieval, input processing, and output generation. The hardware component is connected to a portable personal computer (PC), which in turn is connected to the infrastructure by wireless local area network. The main purpose of Nomadic Radio is managing voice and text-based messages including voicemail, e-mail, calendar entries, news, traffic, and weather updates. The notifications of events are context-sensitive and adapt to the user's prior behaviour. They are scaled dynamically from ambient sound and recorded voice cues to message summaries. The user interacts with the system either by voice commands or tactile input. Nomadic Radio uses spatial audio, with the audio cues positioned in a circle around the user's head according to their time of arrival.

Hubbub: Hubbub [182] is a messenger software supporting social presence and opportunistic social exchange. It runs on mobile as well as stationary clients. In addition to a visual interface, Hubbub relies strongly on non-spatialized auditory output: the users and activities are identified by a unique sound. Changes in the status of a certain user are indicated by first playing the sound cue for "change of status", followed by the sound cue for the specific user who changed their status. In addition to text messages, users can also send each other sound instant messages (SIMs). SIMs are the audio equivalent of emoticons or other commonly used abbreviations, e.g., LOL for "laughing out loud" or BRB for "be right back". SIMs are earcons which can either be sent directly to a user or be integrated into text messages. Unfortunately, Hubbub does not support eyes-free interaction with the system. The

user and SIM selection, as well as text input and output, still require a visual interface and the standard input device of the platform it is running on.

In-vehicle Spatial Auditory Interface: The first auditory interface for interaction with a mobile phone while driving a car was developed by Sodnik et al. [7] All items of the mobile phone's menu and all commands were presented with spatial sounds and played to the driver via six circularly arranged speakers installed in the driving simulator. Each item in the menu was assigned a corresponding sound source. The sound sources were spoken words - readings of the menu items. At each level of the menu hierarchy, the items were placed on a virtual circle which could be rotated around the user's head. A unique, unobtrusive background melody was assigned to each individual branch of the menu. The pitch of the melody was changed according to the current level of the user in the sub-menu. Thus the user could, at all times, maintain an awareness of their absolute position in the menu. Interaction with the system was achieved through a custom made interaction device consisting of a small scrolling wheel and two buttons (left and right), which was attached to the steering wheel.

Mobile Spatial Audio-Conferencing: An interface for navigating between multiple sound streams, such as multiparty phone calls, podcasts, or music, was developed by Dicke et al. [6] The interface uses spatial sound and positions the audio streams on a circle around the user's head. The user can either use gestures with the mobile phone or press keys to interact with the system. Individual streams can be focused on and played in stereo with all other streams muted; they can be positioned on an inner ring with all sources playing at the same time, or they can be pushed away and played from a distance. Panning gestures rotate the ring and allow source selection. Pulling and pushing gestures are used to manipulate the sources – pulling a source toward the user activates it, while pushing it away deactivates the source or pushes it further away. The foreground/background metaphor allows

the user to concentrate on the closer sources; nevertheless, it maintains an awareness of all active streams.

iPod Shuffle: The 3rd and 4th generation iPod Shuffle[16], a mobile music player, uses an auditory interface to navigate between the playlists and to retrieve information about the sound files. The user interacts with the system by pressing buttons integrated into the cable of the earphones. A text-to-speech (TTS) engine reads the playlist or song title, and the user can then either select it or proceed navigating. The iPod Shuffle does not use spatial sound. It is the first commercially available product relying on an auditory interface only.

Auditory Interfaces for Visually Impaired Users

Visually impaired people are permanently unable to use their visual channel for interaction with mobile devices. The tactile interfaces also require a high degree of concentration and are also not very practical in mobile situations. The auditory interfaces on various mobile devices seem to be appropriate and easy to use for various types of applications. Some of them are used for basic navigation and orientation in space, so called electronic travel aids, or ETAs (see [183] for a review), while others enable the normal use of a mobile phone or PDA functionalities, such as browsing through contacts, identifying a caller, reading or writing a text message, etc.

The vOICe: The vOICe is one of the oldest and most interesting projects in this area [184]. It is an attempt to create a navigation tool for visually impaired people based on direct conversion of a video image into an audio signal. The video signal is captured by a normal Web camera, converted into black and white, and then divided into 64×64 individual pieces (pixels). The image is then sonified with an acoustic scanner moving from left to right, column by column. The direction can actually be heard due to the use of spatial effects. One column consists of 64 pixels, and each pixel is presented with a different frequency. The pixels at higher physical positions in the image are coded with high

[16] `www.apple.com/ipodshuffle/`

frequencies, while the low pixels are coded with low frequencies. The intensity of sound of each pixel codes its color intensity from white to black. However, as the system uses an artificial way of coding video information with sound, it requires intensive learning and adaptation by users. The authors reported the system to be highly usable, as it enabled an entirely independent navigation of the visually impaired and helped them see with the use of sound. Initially, the prototype was built on a PC, but, nowadays, it is commercially available for all major types of mobile phones and different operating systems.

Ecological interface design for ETAs: Davies et al. [185] developed an audio based navigation system for blind travellers. They utilized auditory icons to represent objects and their motion paths. For example, footsteps were used to indicate pedestrians moving towards or away from the user. Earcons were used to show the distance of objects. Their pitch was mapped to the object's distance with an increase along a chromatic scale for each shortening of distance and a mapping of playback tempo to the rate of the approach. Larger objects had an initially lower pitch than smaller objects. ILD and ITD were used as localisation cues for object positions.

Museum Guide: A location-aware museum guide on a PDA was another interesting approach designed to help visually impaired users in their orientation and navigation [186]. An advanced electronic guide limited to controlled environments, such as a museum, gives visually impaired users information on artworks or scientific specimens in their original location. The guide is meant to provide a more enjoyable and informative visit. The location of the user is determined and calculated using radio frequency identification technology and an electronic compass. Synthesized speech is used to communicate with users, giving them information on specific artefacts and navigation instructions within the museum.

SYPOLE: SYPOLE is an automatic text reading assistant which runs on a PDA [187]. It scans and recognizes an arbitrary text and reads it to

the user. The input image of the text is captured by a camera and the image is then processed in order to locate the text in the image. An additional layer of large buttons was put on the original PDA's touch screen, as it is almost impossible for the visually impaired users to use any type of touch screen due to their small buttons and relatively high resolution. A speech-based auditory interface is used for output in order to read the text to the user.

Screen Readers for Mobile Phones: In general, screen readers are the basic tool for any visually impaired user, enabling them to read any text on the screen. Screen readers are based on TTS technology, which synthesizes human voice from written text. There are several commercially available screen readers for mobile devices. "Mobile speak"[17] and "Talks"[18] are widely used by Symbian phone users, while "Pocket Hal" and "Smart Hal"[19] are available for Windows-based phones and PDAs.

Conclusion

Computers and other types of electronic devices are used not only in offices and desktop environments but also in various mobile situations such as walking, cycling, driving, etc. Mobile life requires the devices to be smaller and lighter in order to be carried around anywhere and any time. The small displays and keyboards fundamentally change the way we think about user interfaces for these devices and HCI, in general. The audio and auditory interfaces seem to offer an excellent method of exchanging information between a mobile device and a user. Most of the new methods featuring audio are still not integrated into commercially available products (with the exception of iPod Shuffle and special-interest applications), but have matured and proven to be very useful.

Alternatives to the WIMP paradigm, including interaction techniques, input and output modalities, and interface metaphors, are likely to have a

[17] `http://www.visioncue.com/mspeaksmart.html`

[18] `http://www.nuance.com/talks/`

[19] `http://www.yourdolphin.com`

much stronger impact on future mobile devices, in which auditory interfaces may play an import role. Despite this fact, the auditory interfaces will most likely never fully replace the visual interfaces, but rather complement them.

Future mobile devices will not rely on one single interaction or display technique, but will offer a variety of different techniques from which the users can choose, depending on the type of information they are accessing and on their current situation. More products using 3D sound will become commercially available as audio output and computational capabilities of mobile devices increase. Some mobile phones with hardware support for 3D sound can already be found on the market.

2.7 Summary and Discussion

This chapter has presented a review of the related research fields relevant to the research work presented in this thesis. This review began with an introduction to the perception of sound, spatial hearing, auditory memory, and distance perception and a comparison between the sense of hearing and the sense of sight with regard to their applicability in human-computer interfaces. Following this, further factors contributing to the use of spatial sound in interfaces, especially for communication and navigation tasks, such as issues of attention and distraction, were considered, highlighting the main benefits and limitations in each case. Additionally, the discussion presented here reflected upon the role of sound in augmented reality applications and the condition of simulator sickness as a potential confounding factor in the design of auditory interfaces. In the final section of this chapter auditory interfaces and their components and display dimensions were discussed and a review of past and present implementations of sound in interfaces was given. This review revealed several considerations that have to be taken into account when designing a spatial auditory interface for a mobile device:

Spatiality

By adding a second and third dimension to the display, dimensionality can be used to convey addition information. This can either be in form of a (semi-) realistic sound scene rendering where the physical location of a sound

source has a meaning as it might be desired in augmented reality audio[20] or pathfinding and navigation applications[21].

Spatiality can also used in a more abstract and metaphorical manner where position is interpreted in the context of a metaphor, like a clock, and so conveys, for example, the time of arrival of a message[22]. Additionally, distance has previously been used to support multitasking by applying a foreground/background metaphor helping the user to maintain an awareness of multiple streams playing in the background while focussing on one stream playing in the foreground[23]. Given that using spatial sound to communicate absolute distances is a difficult task due to the limitations discussed in section 2.1.5 and the lack of control over the final playback, the granularity that a precise absolute distance display would offer is nevertheless desirable. Therefore, in chapter 3 a new approach towards absolute distance display will be presented and discussed.

It has also been argued that stereophonic and binaural sound have the potential to increase the sense of immersion, presence, and social presence. This can be a valuable effect especially for scenarios in which increased immersion or social presence is desired - social networking applications or games come to mind. As the research regarding the correlation of the recording/playback technique and the subjective perception of presence, social presence, and immersion is somewhat fragmentary, chapter 6 is devoted to further investigating this issue. Based on the findings in chapter 6 the immersive qualities of spatial sound as a means to differentiate between two interface modes are considered in chapter 8.

Simultaneity

Many features of current human-computer interfaces are designed to support both multitasking and *continuous partial attention (CPA)*, a term coined by Linda Stone describing how many users continuously scan their devices for

[20] See [120, 13] for examples.

[21] See [179, 180, 181] for examples

[22] See [168, 14] for examples

[23] See [6] for an example.

news, changes, or updates, briefly focus their attention on some data, and then move on to the next stream of information[24]. In auditory interfaces the continuous presentation of multiple streams of information not only leads to a slowed information retrieval but, due to masking effects, to a complete inability to extract information. While spatially separating individual streams can improve comprehension (30 percent increase in identification scores for an increase of 60 degrees separation), this strategy only works for a very limited number of sources [68]. A partial solution to this problem, aside from limiting the number of simultaneously available streams, is to offer the user a way of prioritizing important streams and neglecting streams of lesser importance. The foreground/background metaphor evaluated in chapter 7 is a viable candidate to help the user focus attention on sources in the foreground, while maintaining an awareness of sources in the background.

While some applications require a continuous playback of streams over a longer period of time, such as a telephone conference application, music and audio book players, or navigation software, other applications may only need a sporadic, highly information enriched "burst", similar to High Density Sonification (HDS) [154]. Overview techniques, for presenting the contents of a folder or a list of currently active items, are a use case for sporadic presentation techniques. As the total playback time scales with the items presented, the form of presentation has to achieve a maximum of item comprehensibility/identifiability with a minimum of total playback time. Chapter 4 presents and discusses such an information presentation approach.

Interaction Technique

Several ways of interacting with auditory or mixed modality displays have been explored. A large number of systems use passive whole body tracking to trigger or activate functions. Indoor movement tracking is usually either realized with infrared transmitters/receivers[25] or with radio system trans-

[24] `http://lindastone.net/2009/11/30/beyond-simple-multi-tasking-continuous-partial-attention/`

[25] For example: [120, 13]

mitter/receivers[26]. GPS is mostly used for outdoor position tracking[27].
For item manipulation, such as selection or changing the volume, or navigation within the system structure, apart from keyboard and mouse input, these four interaction methods or combinations thereof are mostly used:

- Hand gesture tracking[28]
- Device gesture tracking[29]
- Head-tracking[30]
- Speech input[31]

All of these interaction techniques have certain advantages and disadvantages. 3D hand tracking requires either a data glove or some other tracking device that is physically attached to the hand. Alternatively, tracking can be done through vision-based marker or markerless tracking methods, which usually require one or more, often stationary, cameras. For two-dimensional hand or finger tracking, a touch sensitive surface, such as a touch screen, can be used. Device gesture tracking requires the device to be equipped with an appropriate sensor, such as a gyroscope, a digital compass, or an accelerometer. Also, as is the case for hand tracking, the user has to have enough space to perform the gesture. Involuntary movements caused by the environment (e.g. sitting in a car or on a bus) or an activity (e.g. walking), such as vibrations or jerks, can complicate the interaction or create false input.

Head-tracking requires a tracking device being attached to the head. Often such devices are mounted on headphones or caps, helmets, or attached to glasses. For all tracking techniques, initial calibration is required and decalibration, as may be caused by gyroscope drift, are issues.

Speech input requires a microphone. The issues here are privacy and sensitivity against environmental noise, especially other speakers. Usually

[26] For example: [127]

[27] For example: [121, 179, 178, 181]

[28] For example: [162, 16, 167]

[29] For example: [15, 6, 188]

[30] For example: [170, 171, 167]

[31] For example: [14, 123, 189]

considerable effort for system training is necessary to achieve good recognition rates.

All interaction techniques, with the exception of 2D gestures on a touch screen, are prone to issues of social acceptability caused by their novelty and sometimes disruptive nature. Most camera-based tracking is unsuitable for mobile situations, due to changes in lighting conditions and the number of cameras required.

As successful user input is a crucial part of the user interaction design process, chapter 7 explicitly explores the potential of device based gestures for interaction with mobile auditory interfaces.

Metaphors

Interface metaphors not only set the look and feel of the interface, but are a substantial element of the design process. A well chosen metaphor can not only improve the efficiency of the interaction but can also help users understand the functionality of the system and make the interaction more enjoyable [190]. Inconsistencies, on the other hand, can have the opposite effect and lead to confusion and irritation, for example if the system does not behave as expected. When choosing an interface metaphor, good interaction design does not only take into account the context of application and the form factors of the device, but also the social and cultural background of its users, and their prior exposure to and knowledge about other interfaces and the metaphors used therein.

Social Presence

Several applications have taken advantage of the human ability to unconsciously perceive and interpret background sounds to induce a sense of social presence. Cohen [191] and Smith & Hudson [192] have shown how certain "activity indicators" such as the sound of keystrokes, mouse clicks, or even unintelligible crosstalk can improve the sense of social presence in physically disperse cooperative workspaces. Hindus et al. [193] introduced Thunderwire, an "always-on" audio connection within a work group, and gained, among other results, interesting insights into the influences on the group's

self-conception. Hubbub [182] used acoustic awareness cues in a mobile instant messenger to support group awareness and the sense of being connected to colleagues. All these applications not only show how beneficial a sense of social presence and awareness can be to cooperative and social spaces but they also demonstrate the advantages of sound as an unintrusive source of information. The subject of aurally induced perceptions of presence and social presence is dealt with in more detail in chapter 6.

Building on the knowledge and awareness of the current research work presented in this chapter, the next chapters capture the results of my research, which in turn form the basis for the design and prototypical implementation of Foogue, a design concept for eyes-free interfaces for a smartphone, which is described in detail in chapter 8.

Chapter III

Distance in 3D Audio Interfaces

Mapped to a metaphor or a symbolic connotation, distance can add meaning to an object position or its behaviour in human-computer interfaces. For example, if distance is used metaphorically, it could reflect importance, with closer objects having a greater importance than objects which are positioned further away, or distance could be mapped to a time scale with nearer objects being closer in time. The concept could also be used to reflect similarity with a specific distance indicating a specific object type and so forth. Beyond these, the representation of distance can have several other functions: in addition to use in games or other applications concerned with the recreation of a realistic environment or a specific atmosphere, authentic, scaled, or relative object relations are often used in navigation software. Distance is also an important dimension in Electronic Travel Aids (ETAs) for the visually impaired for communicating the arrangement of nearby objects or potential collisions. This chapter addresses the following research questions raised in the introductory chapter of this thesis:

RQ 1.3: How can acoustic distance perception be used as an aspect of interface design?
RQ 1.4: How can acoustic distance perception be improved?

An experimental study of a new method for the display of absolute and relative distance based on acoustics is presented and discussed. This novel approach partially overcomes the difficulties considered in section 2.6.2. For an introduction to spatial hearing, please see section 2.1.2. An overview of the psychophysics of distance perception is given in section 2.1.5.

3.1 Introduction

According to a recent study from the Nielson Company[1], Maps and Navigation Applications are among the top five most popular apps used on smartphones. They are used to find a specific point of interest, to get an overview of an area or, to travel from one point to another, among other things.

For mobile navigation, maps as graphical representations of geographical spaces or locations are often complemented or entirely substituted by verbal information, such as "Turn right in 700 meters." due to the distraction caused by visual interfaces (*eyes-off-the-road*). Several studies have demonstrated that there is an advantage to exclusively auditory or auditory and visual systems over purely visual information presentation [194, 7, 195]. Nevertheless, even auditory voice information systems cause a selective withdrawal of attention that is often referred to as *mind-off-the-road* or *cognitive distraction*. This scales with the complexity of the secondary task [196]. This scaling means that while drivers can cope with the workload generated through step-by-step guidance systems, a much higher impact on the primary task and degraded information extraction from verbally presented information in the secondary task can be expected when the user is presented with more than one chunk of information. This type of information overload would occur if the user wished to know where all gas stations in a radius of 50 km are located, or, if a user in a generic mobile scenario requested *overview* information regarding objects with dimensional attributes. As long as the system knows which specific item the user is interested in, the next way-point, or the nearest gas station, the information can be reduced and presented accordingly without overstraining the user's attentional resources. However, this is not the case with overview information, as the user – and consequently the system – does not know which specific item out of a multitude of items they are interested in. Counterbalancing the quantity of information with an easier and less distracting method of presentation could be a potential solution to this problem. By reducing the amount of information that is verbally encoded and using a realistic 3D sound field to simulate distance

[1] The Nielsen Company (2010): The State Of Mobile Apps: `http://blog.nielsen.com/nielsenwire/wp-content/uploads/2010/09/NielsenMobileAppsWhitepaper.pdf`

information, attentional issues and information overload could be addressed. Unfortunately, realistic sound field synthesis has its own limitations.

While the human ability to localize sound sources is fairly accurate, distance perception can be problematic, due to a range of factors. One of these factors is how we use a sound source's loudness, i.e. the interpretation of the perceived magnitude of its intensity, to judge its distance. The inverse square law predicts sound intensity reduction with increasing distance, but to accurately judge this distance the initial intensity of the source has to be known. Otherwise the listener cannot tell whether the sound they hear is low in intensity because it has travelled over a certain distance and already lost a lot of its intensity, or whether it is fairly near and was never very intense [197]. Even if the listener is familiar with the sound source, intensity based distance cue can be ambiguous, since intensity changes can be caused by both changes in distance and in acoustic power [40]. Naturally, the sound source intensity has to fall between the absolute threshold of hearing and the threshold of pain. Given that human speech has an intensity of around 60 dB when measured at a 1 meter distance, even under ideal listening conditions it will become inaudible when the distance is increased to 1 km (-6 dB per doubling of distance). As the source's intensity cannot be dynamically adjusted to cover a wide range of distances, using a simulation of a realistic sound field would restrict either the displayable distances or the displayable sound sources to a much greater extent than would be acceptable.

To overcome these limitations and to provide an intuitive, eyes-free and scalable solution for displaying objects in three-dimensional space a method is proposed and described in the following section. Its experimental evaluation is included in this chapter as well as a discussion of the results.

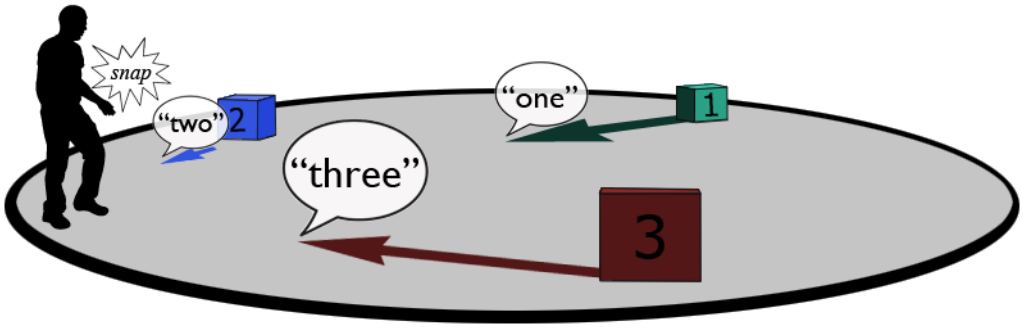


Figure 3.1: The user gives a starting signal and all objects in the environment reply. The travel time of sound is used as distance indicator so that the closest object is heard first, then the second closest and so forth.

3.2 Walkthrough

The method is modeled on echolocation and biosonar techniques used by several animals such as bats or dolphins. By emitting calls and listening to their echoes these creatures are able to gather information about the location, distance, and even type of objects in their environment. To be usable by humans in a user interface, these echolocation techniques were abstracted and simplified. The method is illustrated in figure 3.1: the user emits a sound, e.g. a finger snap, which functions as a starting signal for all objects in the environment to verbally identify themselves. From each object these sound waves spread with the speed of sound (343.2 m/sec) until they arrive at the listener's ears and are heard. In this way distance is coded as travel time or *time-of-flight* into the arrival of each individual sound. In figure 3.1, the object closest to the user, in this case object 2, is heard first, then object 3, and after some time object 1.

In addition, the method utilises spatial sound to add positional information to each signal and thereby the user is not only provided with distance information but also each object's position in three-dimensional space.

3.3 Related Work

Although echolocation is mostly associated with certain animals such as bats (*Chiroptera*) and dolphins (*Delphinidae*), it is not an unfamiliar technique for humans to acquire information about their surroundings. For example,

in 1944 in their work on *facial vision*, Supa et al. [198] found that obstacle detection is possible by stimulation of the auditory system, rather than, as previously hypothesized, stimulation of the skin by air and sound waves.

Stoffregen & Pittenger [199] give a comprehensive overview and introduction to human echolocation and research in the field. Beginning with a review of early experiments from the 1940s onwards, which mostly concentrate on the blind, Stoffregen & Pittenger conclude that both blind and sighted participants have the ability to gather information about an object's distance, and sometimes even its shape and material, through echolocation. Distance judgments are deduced from the time delay between the initial pulse generated by the person and the echo of that pulse. As sound travels at approximately 343 m/sec (c) the delay (t) is the key to the calculation of distance (d). Assuming the pulse is generated at the ears, we have:

$$d = ct/2$$

Although the just-noticeable difference (jnd) for two clicks played successively[2] can be as short as 2 ms, the echo threshold for speech is around 20 ms [29]. When given a choice o which oral sound to use as pulse (e.g. hissing sound or tongue click) participants could even detect very small distances in the range of 30–120 cm with an accuracy of 10 cm. However, distance perception for distances shorter than 2 m is error prone due to the short pulse-to-echo delay (2 m is approximately 5.8 ms).

Schiff & Oldak [200] researched the use of time-to-contact information for judging approach trajectories in relation to an observer. They found that for both visually and auditorily presented information judgments worsen with increasing time to arrival but, that purely acoustical information leads to significantly less accurate judgments when the arrival time exceeds about 4 s.

Most applications of echolocation in human-computer-interfaces are designed to support users with complete vision loss or vision impairments. Applications described in [201, 184, 202] use sonar information for spatial sensing and object imaging, but map the echo information to pitch or rhythm

[2] A click is a sound obtained by applying a DC pulse to a headphone or loudspeaker creating a sound with an abrupt onset and a brief duration.

based metaphors to make them more perceptible to human beings. Users, however, require considerable training to be able to interpret information that is transmitted in this way.

Shiose et al. [203] created a 3D acoustic environment to help blind pedestrians detect passing cars and to offer support for road crossings. The system uses acoustic time-to-contact information in order to calculate the time that will be taken for a car to arrive at the listener's position. They found that, as the speed of the approaching car is increased, both blind and sighted people make fewer accurate estimations of the car's speed, consistently underestimating the time that a car will take to reach them.

With an intent similar to [203], Davies et al. [185] propose an interface for blind users that can be used to display travel information gathered by either sonar or video devices. As with most of the projects mentioned above, Davies et al. match distance, size, trajectory, and velocity information to properties of sound. For example, changes in the amplitude of the earcons used for the task indicate distance, changes in pitch to indicate size, and changes in tempo for velocity. Besides earcons they used auditory icons to represent the type of object and/or its motion, such as footsteps representing the motion of other pedestrians. Unfortunately, a formal user study was not conducted and there is no information on the usability of the approach.

Talbot et al. [204] make suggestions for procedures to improve the display of spatial information, especially distance, addressed mainly but not exclusively at blind users. They evaluate *ecological cues*, namely intensity, spectral filtering, and the ratio of direct to reverberant energy, and *non-ecological cues*, such as pitch, temporal variation, and beat rate. Their results show that *ecological cues* consistently yield the best results both in response time and error rate. However, they point out that intensity loss over longer distances exposes a practical problem; distant objects may become imperceptible due to too much intensity loss or masking effects from high ambient sound levels.

Previous approaches that use echolocation in human computer interfaces are scarce, often contextually bound to ETAs, and do not use the raw echolocation data but mediate them, i.e. the information gathered through echolocation is mapped to a second set of sounds, usually earcons, and their properties. This seems to be a practicable approach, especially for a real-time

display of multi-component environments, such as rooms, streets, or crossings. However, there has not been an effort to tap the full potential of the human ability for echolocation in an user interface.

The study presented in the following section is an initial investigation into the efficiency of one component of echolocation, i.e. the travel time of sound. Other factors, such as distance-dependant changes in frequency, amplitude loss, and reverberation have been neglected to allow for an non-confounded view on the potential of travel time for displaying distance information.

3.4 User Study: Simplified Echolocation vs Verbal Information

The purpose of this study was to verify whether a) the proposed method modeled on simplified echolocation indeed enables users to conceive the distance of objects and b) if so, with what accuracy. Furthermore, it was of interest to learn more about the method's strength and weaknesses in different scenarios when compared to traditional verbally coded positional information ("Object One, North, 400 meters"). For the sound design spoken identifiers were preferred over auditory icons and earcons for several reasons:

- Sounds the user is familiar with, such as speech, can facilitate distance localization [197, 40]
- Earcons and auditory icons require a training phase and are prone to cause confounds due to possible memory effects and decoding errors
- Spoken identifiers allow for a higher comparability to the "verbal" condition than non-speech audio
- Many users are familiar with receiving verbal distance and directional information from either other humans or GPS based in-vehicle navigation software

3.4.1 Tasks

Tasks used in this study were designed to gain a general understanding of participants' ability to understand and intuitively decode travel time of sound as a distance indicator. Furthermore, the extent to which each method, verbal information (referred to as *verbal condition*) and travel time (referred

to as *bat condition*) impacts task performance for the following three tasks was investigated:

- Task 1: Participants were asked to estimate the distance of one object with only one object being displayed.
- Task 2: Participants were asked to name the object closest to them out of six objects displayed.
- Task 3: Analogous to the "closest" task from Baudisch & Rosenholtz [205], participants were asked to name the object that is closest to a given target object (figure 3.2).

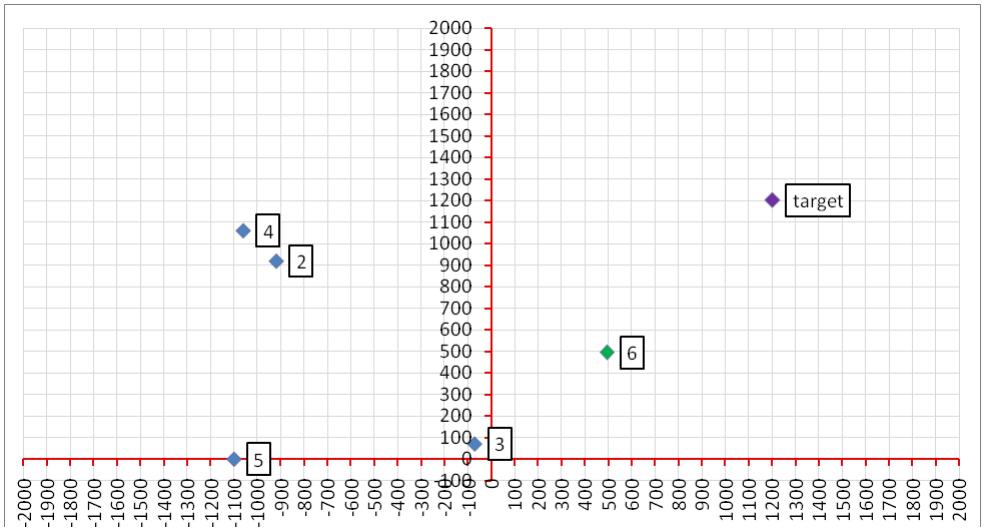


Figure 3.2: Visualization of a typical configuration for task 3. In this case object 6 is closest to the target.

3.4.2 Apparatus Interface

Participants used the visual interface depicted in figure 3.3 to start the next trial or to enter their distance estimation in meters for task 1 or object number (1, 2, 3, 4, 5, or 6) for tasks 2 and 3.

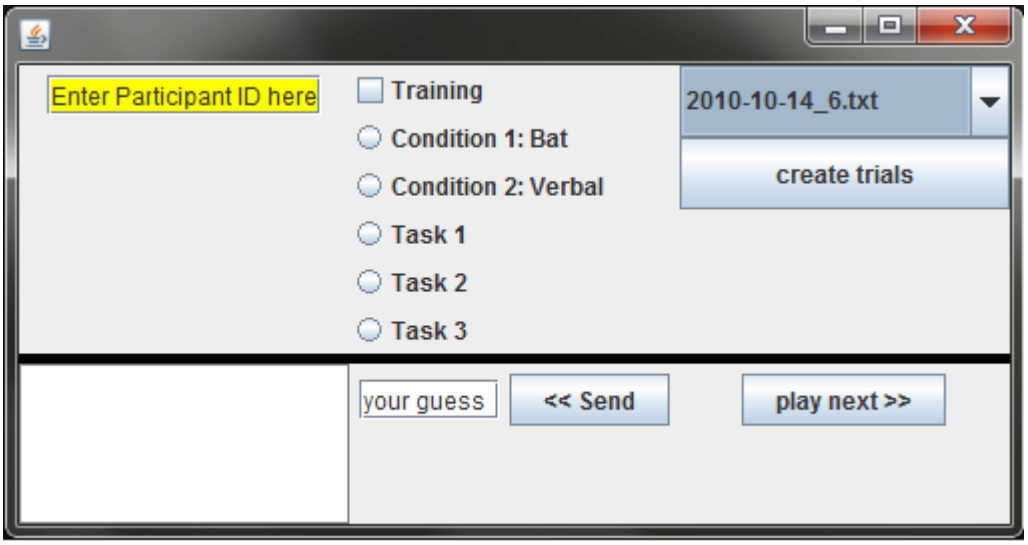


Figure 3.3: Interface used by participants to enter distances or object names and to start the next trial.

3.4.3 Experimental Design

The study design was a within subjects 2×3 (Location Technique×Task) design with 30 repetitions for each cell. Table 3.1 shows all distances and table 3.2 all angular directions used in the experiment. For task 1 the deviation in the participants' estimation of distance from the actual distance was logged. For task 2 and 3 the error rate, i.e., number of times the participant chose the wrong object, was recorded.

Interface order, distances, and positions were counterbalanced. Participants received 10 minutes of training at the beginning of each task. Participants were aware of the distances used in the experiment. The study took approximately 70-80 minutes per participant. After the study participants were debriefed and compensated with a cinema ticket.

3.4.4 Hypotheses

For task 1, H1: The verbal condition will outperform the bat condition. The verbal condition will have a mean error rate close to zero as participants can simply write down the distance information (e.g. 500 meters) in the

Distance (meters)
100
300
500
700
900
1100
1300
1500
1700
1900

Table 3.1: Distances used in the experiment.

Direction(degrees)	**Direction (cardinal)**
0	north
45	north-east
90	east
270	west
315	north-west

Table 3.2: Horizontal positions used in the experiment.

given string "One, North, 500". In contrast, in the bat condition, participants have to decode the metaphor and estimate the distance which is much more error-prone.

For task 2, H1: The bat condition will outperform the verbal condition. The bat condition will have an error rate close to zero as the first sound played is always the closest.

In the case of the verbal condition, participants have to listen to all six sounds and memorize the distance information of either all objects or only the object that is currently closest. As this is inherently a memory task (with 6 objects compared to just one object in the bat condition) it presumably produces a higher cognitive load with lapses in concentration or distractions resulting in more errors.

For task 3, H1: The bat condition will produce a lower error rate than the verbal condition. The verbal condition requires a conversion of the given verbal information into a 2D (visual) mental map. This is a process in which the information needs to be decoded, objects are to be placed on a mental map, and distances need to be calculated. In the bat condition the extraction of information is intuitive, as the sound itself already carries that information.

3.4.5 Apparatus

The experiment was run on a Windows 7 PC with an Intel Core2 duo (2.4 GHz, 2.4 GHz) processor with a standard mouse and a qwerty keyboard. A Creative SB X-Fi Sound Card was used and the spatial sound rendering was achieved using the 3D sound library OpenAL[3]. Objects' handlers, and distance and direction information for the verbal condition, were generated using Apple's Alex voice for Leopard[4]. During the experiment participants wore Sennheiser HD 555 stereo headphones.

3.4.6 Participants

Thirteen randomly recruited participants volunteered for the experiment ranging in age from 22 years to 36 years (M = 26.3 years). Eleven participants were male, two female. Four participants had musical training and none reported hearing problems.

3.5 Results

3.5.1 Task 1 - estimate single object distance

As the data were measured on an interval scale and were not normally distributed a Wilcoxon Signed-ranks test was conducted to evaluate whether or not the representation of information had an influence on participants ability

[3] `http://connect.creativelabs.com/openal/`

[4] `http://www.apple.com/accessibility/macosx/vision.html`

to judge the distance of an object. Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the data set. The results indicate a significant difference with Z = -3.06 and p = .002. The mean of the ranks for the verbal condition is 0, while the mean of the ranks for the bat condition is 6.5 (N = 12).

Participants misjudged the distance of the target object in the bat condition by an average of 149 meters, while there was only one misjudgment in one single trial for the verbal condition. As illustrated in figure 3.4, the minimal misjudgment for the bat condition lies at 86.67 meters and the maximum lies at 226.67 meters.

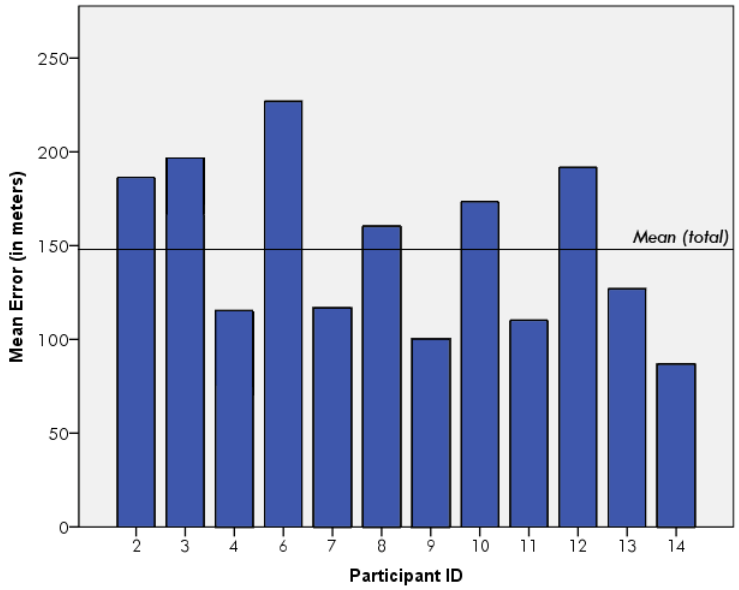


Figure 3.4: Task 1: Average misjudgments of distances in the echo condition by participant.

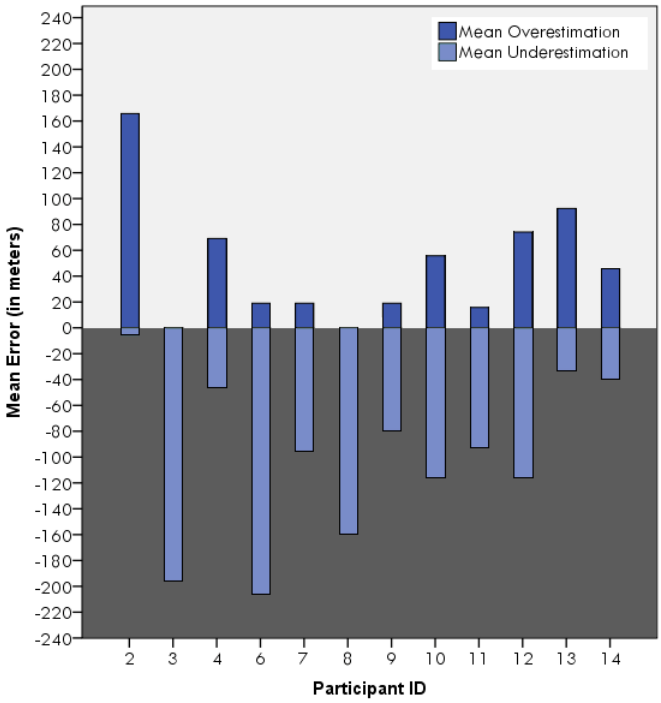


Figure 3.5: Task 1: Average misjudgements by participant in the echo condition split into overestimations and underestimations.

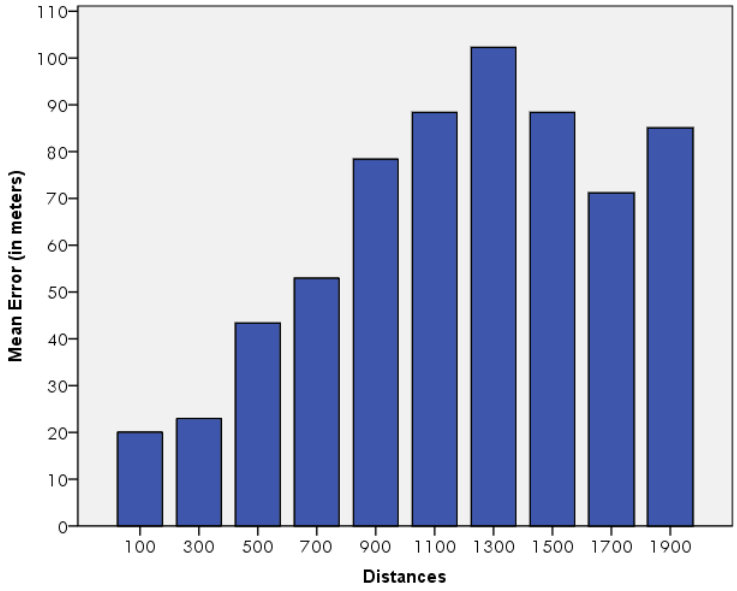


Figure 3.6: Task 1: Differences in mean error rate by object difference in the echo condition.

Figure 3.5 depicts the tendency of participants to underestimate the distance of the object. The general misjudgment of 149 meters is split into 1/3 overestimation and 2/3 underestimation of the object's distance. While the cumulative mean of the overestimation is 49.5 meters, the cumulative mean of the underestimation is almost twice as much at 99.5 meters.

Figure 3.6 implies a linear increase of misjudgments until an object distance of 1300 meters. This figure would be in line with prior findings [39, 197, 40, 203]. To verify a possible relation between the objects' distance and the extent of the misjudgement, and as the data was not normally distributed and ordinally scaled, a Kruskal-Wallis Test was conducted. It indicates a significant difference in the medians with $\chi^2(9, N = 30) = 26$, $p = .002$. Since the overall test is significant, follow-up Mann-Whitney Tests were conducted to evaluate pairwise differences among the groups, controlling for Type I error across tests by using the Bonferroni approach. The results of these tests indicate a significant difference between the shorter distances (100, 300 meters) and the longer distances (1100, 1300, 1500 meters).

3.5.2 Task 2 - closest of six objects

Error rates were generally low, indicating that both methods were suited to identify the object closest to the listener. As the data was not normally distributed and, as the data had been measured on an interval scale, a Wilcoxon Signed-ranks test was chosen to investigate on possible differences between the groups. It showed a significant difference with Z = -2.59 and p = .008 (N = 13) between the bat and the verbal condition. While there were no errors made in the bat condition (mean of the ranks is 0), a total of 14 errors occurred in the verbal condition (mean of the ranks is 4.5). The result indicates that in the verbal condition significantly more errors were made in identifying the object closest to the listener. On average 3.6 percent errors were made with a maximum of 13 percent (4 false identifications). Figure 3.7 shows the mean error rate per participant (left) and the mean error rate across all participants (right).

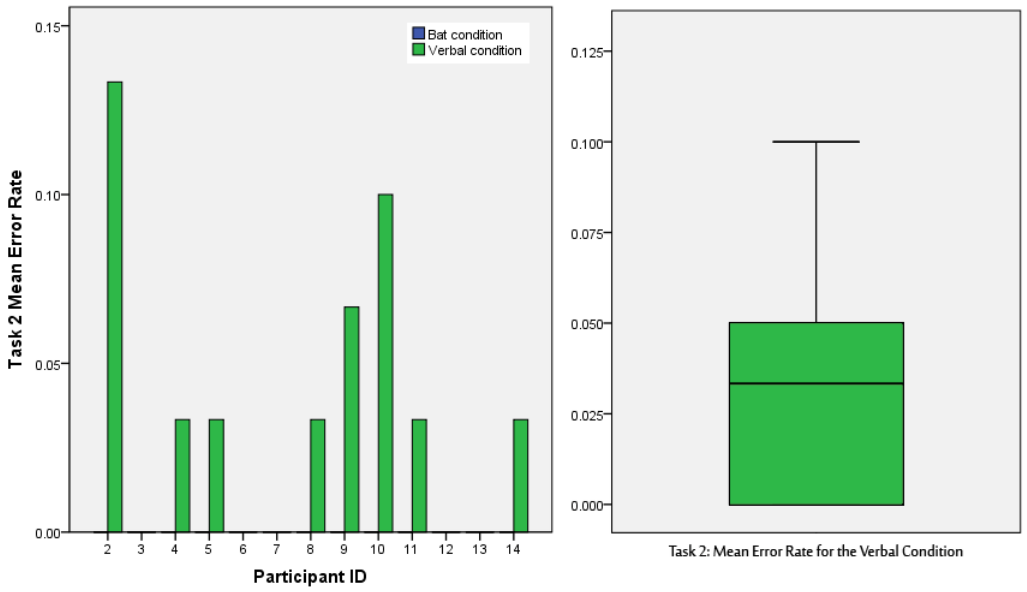


Figure 3.7: Task 2: Average error rate by participant (left) and total mean error rate across all participants (right). 0.10 error rate equals 10 percent errors.

3.5.3 Task 3 - object closest to target

No significant difference between the two conditions was found for the third task. Mean error rates were fairly high with 52.4 percent in the bat condition and 46.5 percent in the verbal condition (see figure 3.8).

Figure 3.9 gives an overview of individual participants' performances. Participants 2, 4, 5, 7 and 10 showed a clearly elevated error rate for the bat condition while participants 6, 9 and 14 have moderate to strong elevations in error rates for the verbal condition.

A post study interview revealed an overall preference (11 of 14) for the bat condition. Participants found the applied method to be faster, more intuitive, and less exhausting than the verbal method.

Task 3 - Observations

Certain arrangements of sources lead to either elevated error rates in the bat or in the verbal condition. Figure 3.10 shows an illustration of a typical

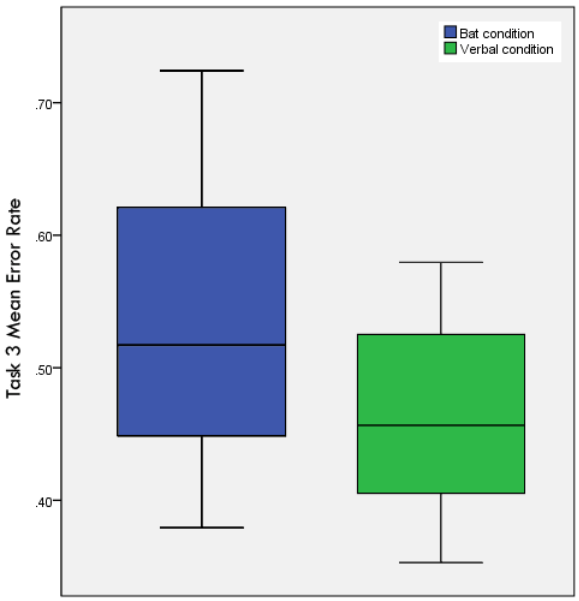


Figure 3.8: Task 3: Mean error rates for the bat and the verbal condition across all participants. .50 error rate equals 50 percent errors.

error-prone setup for the bat condition. Although it is visually obvious that object ② is closest to the target, 11 participants chose object ❺ in the bat condition. As the bat condition works on the basis of the travel time of sound waves between the object and the listener, all sound sources are played in order of their distances from the listener (not from the target object). In this case the target object was played first (1500 meters distance to listener), then object ❺ (1700 meters), and then object ② (1900 meters). Although object ②'s distance to the target is only 400 meters, while object ❺'s distance is 1238 meters, playback order suggests that object ❺ is closer to the target.

Figure 3.11 is an example configuration of an error-prone setup in the verbal condition. To reduce the workload participants stated that they tried to exclude objects at remote angular positions (in this case objects on "west" and "east" positions) as participants made the presumption that these were not the closest objects. But in the example shown in figure 3.11 five out of six objects are on the same azimuthal position ("North") and only one object can be excluded. As all objects are played in a random order, it is challenging for the participant to memorize all objects' positions to memorize

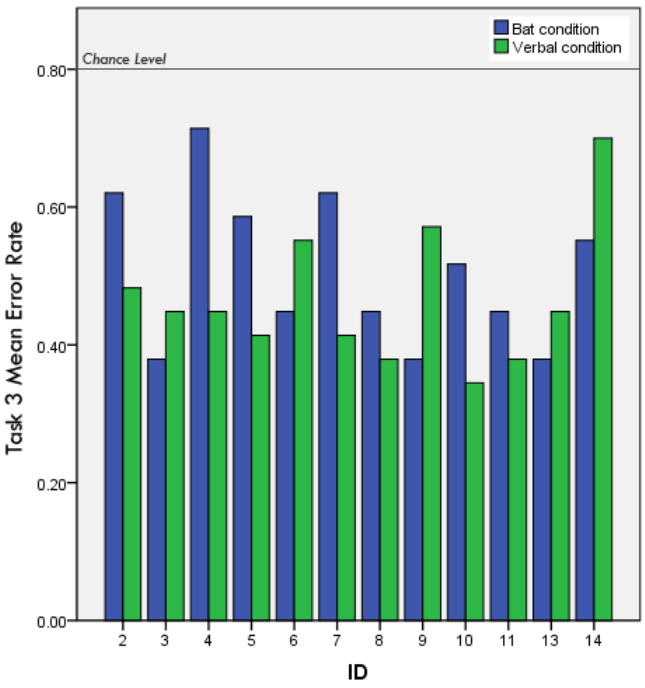


Figure 3.9: Task 3: Mean error rates for the bat and the verbal condition by participant. 0.40 error rate equals 40 percent errors.

the positions of all objects and match them with the target's position.

3.6 Discussion & Conclusion

The results confirm the first hypothesis: for tasks requiring a user to have a precise understanding of an object's distance the verbal condition performs best (task 1).

For tasks in which the user seeks to find out which object of several is closest by (or farthest) the bat metaphor is most suitable (task 2), which confirms the second hypothesis.

The third hypothesis could not be confirmed as no statistical difference between the conditions – for the given experimental setup – was found. The analysis of participants' performance in task 3 and the post study interviews revealed that certain factors had stronger impacts than had initially been suspected. It seems that when in doubt, participants assumed spatial

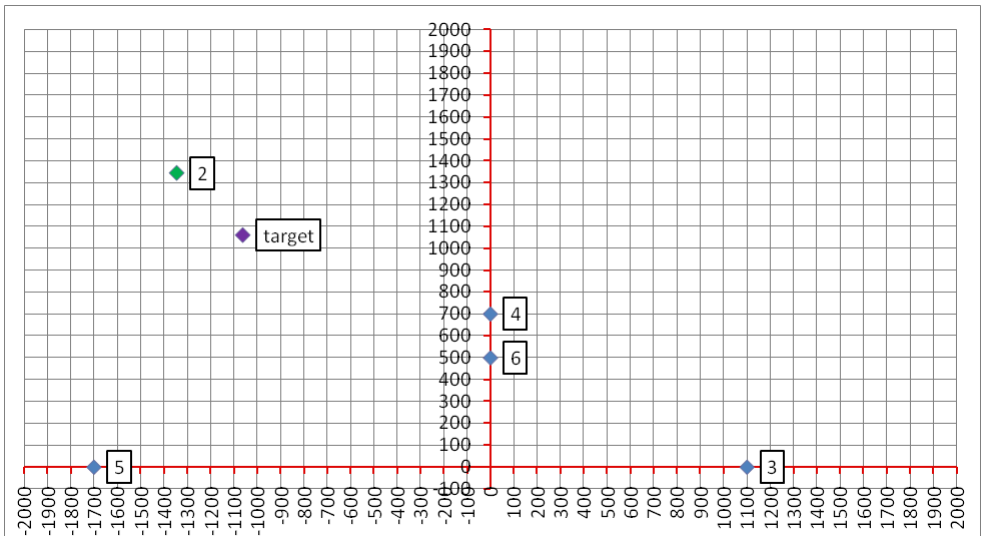


Figure 3.10: Example of *distance to target* vs. *distance to user* confusion for the bat condition in task 3.

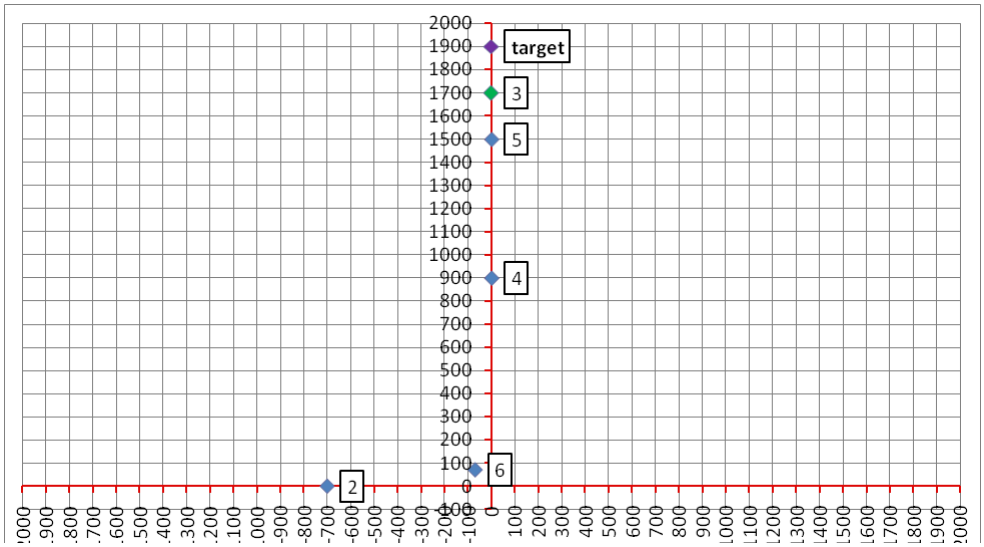


Figure 3.11: Example of a setup in task 3 that lead to a high error rate in the verbal condition (and low error rate in bat condition).

proximity from temporally close sources in the bat condition and tended to neglect angular differences. The Gestalt Theory's Law of Proximity [206] suggests that temporal proximity may induce the mind to perceive two or more temporally close objects as spatially grouped. Figure 3.12 shows the effect in a schematic diagram with C and B having the same distance (and travel time) to the listener. They are played at the same time and hence appear to be spatially close although they are located on opposite azimuth positions and A is in fact closest to B. Also, neglecting angular information in

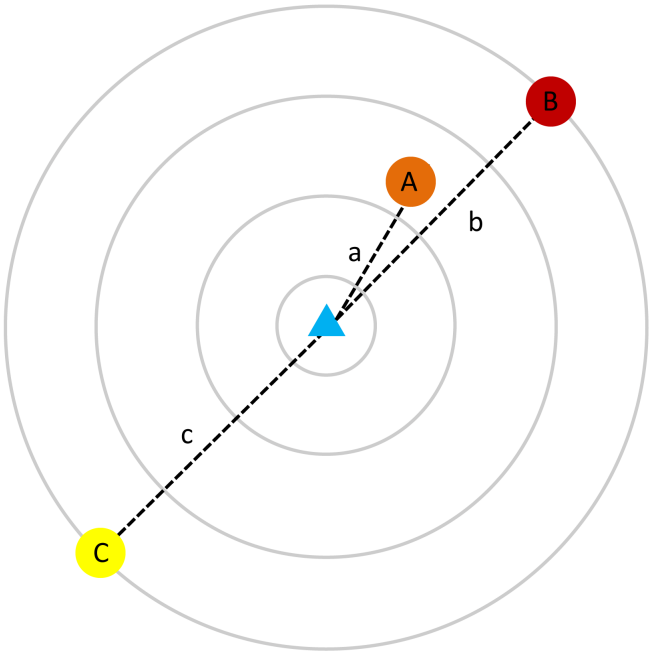


Figure 3.12: Schematic view of the *temporal proximity* vs *spatial proximity* problem. While the length of c and b are almost the same and therefore they are played in short temporal succession, A is notably closer to B.

the bat condition may have been an attempt to reduce complexity similar to focusing on objects on the same axis as the target's in the verbal condition.

Another counterintuitive factor that may contribute to the error rate of the bat condition in task 3 is that object-to-target distance scales with the distance from the listener. As illustrated in figure 3.13, objects C and B have the same distances (c and b) to the listener and are therefore played simulta-

neously. While C' and B' also have the same distance to the listener and are also played simultaneously, the distance (a') between C' and B' has increased compared to distance (a) between C and B. Therefore, when deciding which object is closest to a target object, the listener has to take into account that the differences in travel time of individual sounds cannot be linearly mapped to the distances between objects.

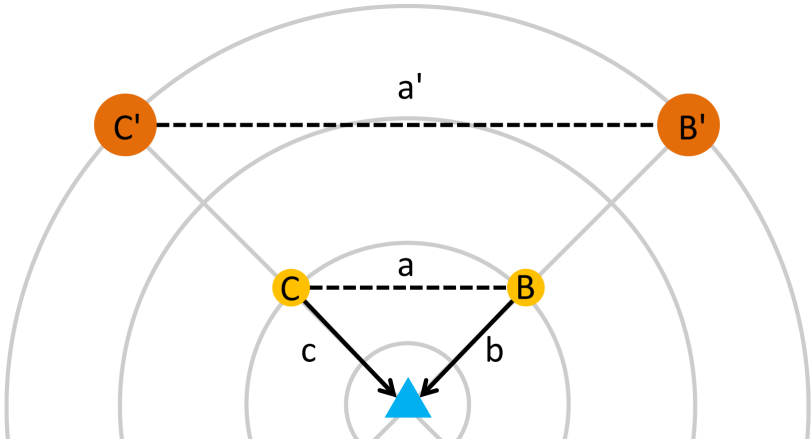


Figure 3.13: Illustration of the increase in object-to-object distances with increase in distance from the listener (triangle).

This chapter has addressed two of the superordinate research questions raised in the introduction to this thesis. **RQ 1.3:** How can acoustic distance perception be used as an aspect of interface design? and **RQ 1.4:** How can acoustic distance perception be improved? When applying either the bat or the verbal method in a user interface the following individual features should be considered:

- The travel time of sound, as used in this study, is only applicable for distances within a certain range. As the distance is mapped to the actual speed of sound (343.2 m/sec), displaying objects that are more than 3 km is likely to be tedious for a listener as well as rather inefficient because it would require focussed attention over a period of 8.74 seconds or more, which is at and beyond the limit of the attention

span [207]. To display bigger ranges the method would have to be adjusted, for example by introducing a scaling factor (zoom) as it is common in visual displays of maps.

- When displaying the distances of multiple objects by travel time the total duration of the playback will depend not on the number of objects but on the distance of the farthest object. Therefore, for the representation of a large number of objects the bat metaphor is significantly less time consuming than the verbal condition. The efficiency of using travel time is defined by the distance range; the efficiency of the verbal method is defined by the total number of objects, and decreases significantly beyond 6-8 objects.
- The correct localization of each individual sound source is crucial for the bat method. Spatial sound representation requires either a loudspeaker arrangement that is capable of simulating a spatial sound field or that the listener has to wear headphones and listen to a simulation. In most mobile scenarios a fixed loudspeaker setup is very difficult to realise, with the exception of using the onboard entertainment system in a vehicle. The verbal condition, on the other hand, only requires monophonic sound and is therefore less dependent on specific hardware and software. However, even for monophonic playback at least one speaker is required unless headphones are worn.
- Participants were fairly accustomed to being given verbally coded positional information from their everyday life, while the bat method was completely new to them. Training and more practice may have a stronger positive impact on the bat method's performance than on the verbal method's performance.

The insights gained in this chapter contributed to the design of the eyes-free interface prototypes described in chapters 7 and 8. While in this chapter the potential for communicating distance through an echo metaphor was explored, the question of how sound can be efficiently used to obtain an overview of items is revisited in chapter 4. The results from the following chapter 4 can in turn be used to further specify the limits of the echo method. In summary: the best compromise between efficiency and effectiveness, i.e.

playing a number of sounds in the shortest time possible and good item detection and comprehension, can be achieved by leaving a 200 ms interval between the onsets of the sounds. This means that if the echo metaphor is used to display objects, the inter-object distance on the horizontal axis should not fall below approximately 68.6 meters.

Chapter IV

Item Detection and Overview Information in Lists

4.1 Introduction

The study presented in this chapter investigates the viability of auditory representation of information in a scenario where the user wishes to gain an overview of available items. The effect of various design parameters of auditory information display on user performance is measured in two basic information retrieval tasks: detecting one specific aurally presented item among a set of items and, estimating the relative number of instances of a given item in two sets of items. The following research questions are addressed in this chapter:
RQ 1.1: How can spatial sound be utilised in an eyes-free interface?
RQ 1.2: What are the advantages and disadvantages of using spatial sound compared to stereophonic or monophonic sound?
RQ 3: What is a good way to help users obtain an overview of available items and options?

In this chapter a usage scenario is assumed in which objects surrounding the user are presented via sound. The user may be interested in getting an overview of the surrounding objects, or in determining the presence of a target object. The amount of information that can be presented by means of audio is limited by its sequential nature but the advantages of audio include that it does not interfere with the visual modality or require a line of sight. While this increases the number of simultaneously displayable sounds, it also reduces their audibility and comprehensibility due to masking effects and auditory memory limitations. Therefore, to derive practical guidelines for auditory interface designers, the study is designed to identify those design criteria, that contribute most to low error rates while the overall playback

time is sought to be minimized while efforts are made to minimize the overall playback time. The criteria investigated are:

- Interstimulus onset intervals (ISOI): 50[1], 100, 200, and 400 ms
- Encoding strategy: synthesised speech versus earcons
- Sound reproduction method: spatialised audio using a multichannel loudspeaker system versus diotic playback via headphones

The structure of this chapter is as follows: Section 4.2 presents an overview of prior research directly related to this study. Section 4.3 introduces the experimental setup and procedure, while section 4.4 summarizes the results of the user study. A discussion of the results can be found in 4.5. Section 4.6 concludes the chapter.

4.2 Related Work

An extensive and well-established body of research has identified the psychoacoustic phenomena contributing to how we perceive different sound types in various constellations and qualities. Most of the research has been conducted in adjacent disciplines and is only partially applicable to interface design. For an overview of literature dealing with parallel source or stream presentation, please refer to sections 2.6.3 and 2.1.4. A short introduction to the limitations of auditory memory is given in section 2.1.3.

Directly related to the study presented in this chapter are findings deriving from research conducted by Massaro [208] that indicate improved identification performance for pure tones varying in frequency as interstimulus onset intervals increase from about 40 ms to about 250 ms. Lorho et al. [209] studied listeners' abilities to segregate spatially separated earcons[2]. They compared ISOI of 0, 0.5, 1, and 2 seconds and found best performances for localisation accuracy, response time, and error rate for 0.5 and 1 second delays.

[1] 50 ms ISOIs were only used in task 1. The decision not to examine an ISOI of 50 ms was based on the finding of a pre-test in which participants had severe problems identifying the *key* items when the ISOI was as short as 50 ms.

[2] Please see section 2.6.1 for an introduction to earcons.

However, so far no studies have been conducted that methodically cover all contributing factors and from which guidelines to improve the design of auditory interfaces can be derived.

4.3 User Study

The goal of this study was to gain insights into the efficiency of auditory display techniques for item detection and overview information of items in a list or - if spatialised - in a designated area. The test simulated two basic information retrieval tasks: detecting a *key* sample among a set of distractor samples and estimating the relative quantity of *key* sample instances in two sets. The sample sets were presented to the users as a list of pre-recorded sound samples staggered with different onset intervals.

For both tasks items were presented to the test subjects as a list of pre-recorded sound samples. Each sample represented an item in the list. The tasks were performed under a number of test conditions with differing design parameters: the playback setup and sound type used to display the item, the number of sound samples presented to the user, and the onset delay between them. The effect of these parameters on the user performance is derived from the error rates for completing each task under the various test conditions. During the listening test each test subject completed both tasks.

Task 1

Figure 4.1 gives an overview of parameters for task 1. Test subjects were presented with a list consisting of 15 items. For each list, they were asked to determine the presence or absence of the sound sample representing the *key* item. Pure guesswork would result in an error rate of 50 percent. The motivation behind this task was to determine the ability of test subjects to notice a certain *key* sound sample among a sequence of distractor samples. For the multichannel loudspeaker playback, test subjects were asked to state which loudspeaker they thought the *key* sample was being played from. For the diotic headphones playback this subtask was omitted, since the samples were not separated spatially.

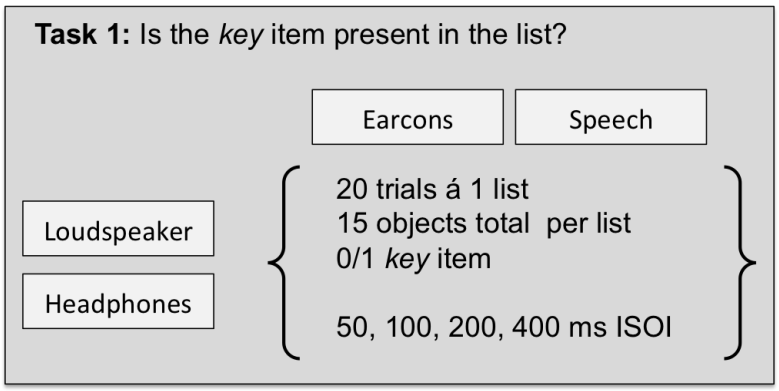


Figure 4.1: Task 1 - Conditions.

Task 2

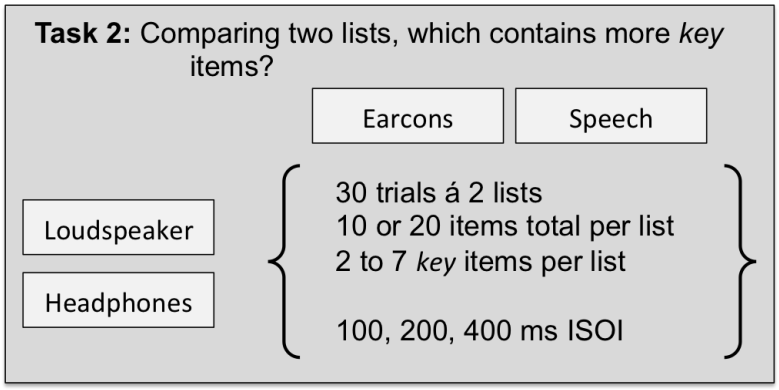


Figure 4.2: Task 2 - Conditions.

Figure 4.2 gives an overview of parameters for task 2. In this task participants were asked to determine, which of two lists contained more *key* items or whether they think the lists contain the same number of objects. Each list contained between 2 and 7 *key* items out of 10 or 20 items in total. Subjects were asked to state whether the occurrence of *key* items in each scene was the same, or, in case it differed, to identify which list contained the greater number of items". Thus, pure guesswork would result in an error rate of 66.7 percent. Participants were not informed that none of the cases

presented to them contained lists with the same numbers of *key* items.

Each item in the list was represented by one sound sample. The encoded samples of each list were concatenated to sequences by staggering the onset times of the samples, resulting in an onset delay between consecutive samples. Varying the onset delay affects the overall duration of the playback. A small onset delay speeds up the playback, but increases the overlap between sound samples. In the study, onset delays of 50 (only for task 1), 100, 200, and 400 ms were compared in terms of their effect on the user performance. This range of onset delays was found to be suitable for the given tasks and sound samples in informal pilot tests.

A central design parameter of auditory displays is the encoding strategy. In this study, synthesised speech and earcons were compared in terms of the user performance. Both are established for the use in auditory display due to their flexibility and performance. The choice of these sound types was based on the assumption that the information to be displayed would be textual and abstract. This ensures that the findings can be generalised to various application areas (for example options in a menu, files contained in a folder, marks on maps or geotags), the approach also rules out alarms, auditory icons, and sonification as potential rendering techniques. Furthermore, text-to-speech synthesis supports the automated encoding of textual information, making it an attractive option for a wide range of applications.

In terms of the audio playback, two different set-ups were compared: spatialised audio using a multichannel loudspeaker system and diotic playback via headphones. As illustrated in figure 4.3 12 spatial locations were simulated for the loudspeaker conditions. Figure 4.4 shows a picture of the test environment including some of the loudspeakers.

The loudspeaker playback condition represents the "gold standard" for spatially spread out lists used, for example, in three-dimensional audio applications: Its spatial reproduction is optimal, since it is implicitly processed by the binaural room impulse response of the user. In a mobile context, headphones playback is more practical, since it does not compromise portability. However, the reproduction of spatial sound over headphones remains a challenge. Accurate externalisation and front-back confusions are common problems related to headphone playback [29]. In the listening test, non-

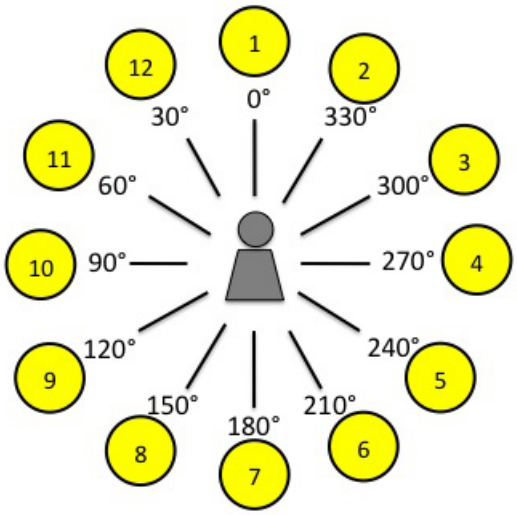


Figure 4.3: Loudspeaker layout for both tasks.

spatial playback was implemented as an alternative to spatial loudspeaker playback, by presenting the audio AR content diotically to both ears of the listener via headphones.

Hypotheses

H1: It was hypothesised that both tasks would generate a lower error rate with longer ISOI than with shorter ISOI. As longer inter stimulus intervals reduces the overlapping of sound samples the occurrence of masking effects is minimized.
H2: The pre-study suggested that for both tasks speech is easier to identify than earcons, although this depends upon:

- The word length and therefore the potential for overlapping and masking effects
- The uniqueness of both acoustic and semantic characteristics of the target item(s)
- The distinctiveness of the earcon design

H3: For task 1, given the results from the pre-study, the detection of the *key* object's position is hypothesised to be "hardcoded" to its identification, i.e.

Figure 4.4: Picture showing a participant in the loudspeaker condition during the training phase. The mounting of loudspeakers is highlighted to show its correspondence with the illustrations in figure 4.3.

the moment a sound object is noticed, its position is obvious to the participants. Very low error rates are expected for the localisation sub task.
H4: For task 2, it was hypothesised that small differences between the number of *key* items in both scenes would be more difficult to identify correctly than large differences. The goal of task 2 was to prove this hypothesis, and to find the detection threshold for the relative difference of *key* items in the two lists.

4.3.1 Apparatus

The study was conducted in a multi-purpose research space with a reverberation time of about 300 ms. For the loudspeaker playback, sound samples were reproduced through a multichannel loudspeaker setup, consisting of 12 Genelec 1029A loudspeakers arranged in a circle of radius 5 m at 30 degree steps as illustrated in figure 4.3. During playback, each sound sample was randomly assigned to one loudspeaker, thus randomising the direction of each sample on the horizontal plane. This is equivalent to randomising the positions of items in the list. For the headphone playback, Sennheiser

HD 212 Pro and AKG K66 headphones were used. The sound samples were played back diotically, at equal loudness levels, and without reverberation or spatialisation.

4.3.2 Sound Samples

Both earcons and synthesized English words were used. Six highly distinct earcons with unique rhythms, melodies, and MIDI instruments were designed. The instruments are listed in table 4.1. Each earcon had a duration of 1 second. To ensure equal loudness, the signal levels of all samples were normalised using A-weighting. No earcon "grammar" or hierarchy was employed in the study.

Object Name	Earcon: Instrument used
Book	Bells
Chair	Bass
Couch	Drums
Cup	Guitar
Keys	Saxophone
Microwave	Whistle

Table 4.1: Items used in the study and the MIDI instruments used for the earcon design.

The synthesized single word speech samples used varied in duration from 400 ms to 800 ms depending on the word. The speech samples were generated by using the Mac OS X inbuilt "Alex" voice. The words that were utilised during the test are listed in table 4.1.

4.3.3 Procedure

The study design was a fully randomised within-subject design. 22 subjects (6 female), aged 19 to 43 years, participated in the study. None of the participants reported any hearing impairments. The duration of the whole study was about 90 minutes per subject. Upon completion, each test subject was debriefed, compensated with a cinema voucher, and dismissed.

The study was organised into two sets of four rounds, each consisting of a different combination of task, playback setup, and sound type. To minimise learning effects, the order of the test rounds was randomised. At the beginning of each round, the test subjects were introduced to the task and the sound sample representing the *key* item. As an example, in task 1, test subjects would be presented with the *key* item (e.g. a synthesised speech sample or earcon representing the item *couch*), and then asked to detect whether that *key* item was present in the list or not. In task 2, the tests subjects would be asked to determine which of two lists contained more *key* items (i.e., in analogy to the previous example, which list contains more instances of the *couch* sample). One test round consisted of 20 (task 1) or 30 (task 2) different trials. Each list was played back as a sequence of speech or earcon samples, through loudspeakers or headphones, depending on the test condition. The range of these test parameters was chosen based on similar studies and a pilot study.

4.4 Results

Each response of the test participants was categorised as either "correct" or "incorrect". For each tested condition, the total error count was calculated. The error count data were summarised in a contingency table, with columns representing different test categories. Pairwise Pearson's chi-squared tests were performed on adjacent columns of the table to test the null hypothesis that there is no association between the test category and the error count. If more than two columns were compared, the Holm-Bonferroni correction was applied to p-values.

4.4.1 Task 1 – Is the key item present in the list?

The *key* sample was present in 80 percent of all trial sets. Because test subjects were unaware of the correct answer distribution, guessing whether or not the sample was present would result in an error rate of 50 percent. Table 4.2 summarizes the error rates for task 1. The results show that error rates decrease with increasing onset delay between sound samples (see figure 4.5). A chi-squared contingency table test of error count data revealed

the effect to be significant with $\chi^2 = (3, N = 79) = 119.35$, $p = .031$. Pairwise comparisons of adjacent onset delays with Holm-Bonferroni correction [210] show a significant decrease of the error rate for each increase of the onset delay.

	ISOI [ms]				**Playback type**		**Sound type**	
	50	100	200	400	Headphones	Loudspeaker	Earcons	Speech
Trials	440	440	440	440	880	880	880	880
Errors	104	46	26	11	80	107	73	114
Errors [%]	24	10	6	2	9	12	8	13
p	<.001	.037	.037		.044		.002	
V	.175	.083	.085		.050		.076	

Table 4.2: Task 1: Error count table with results from Chi-squared tests indicating statistically significant differences between all adjacent columns of the independent variables. V denotes Cramér's V.

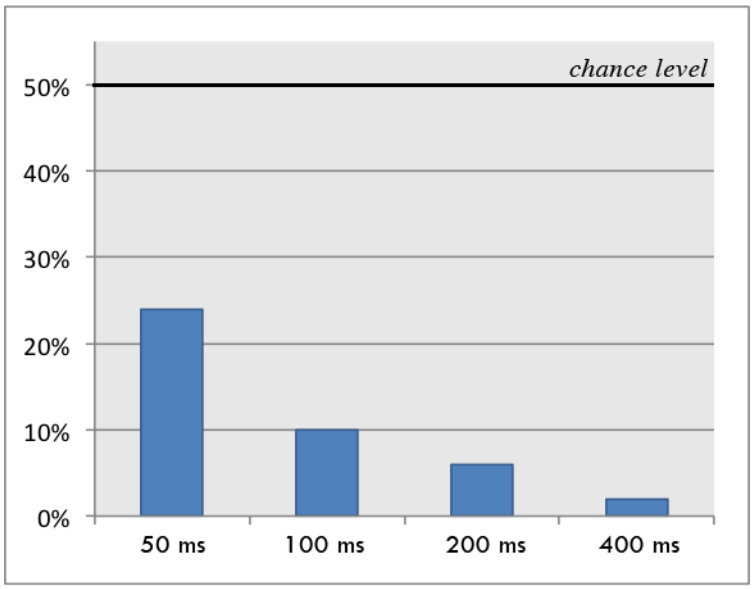


Figure 4.5: Task 1: Pecentage of errors by ISOI.

The average error rate as a function of the playback condition differs by 3 percent, with the diotic headphones playback yielding slightly better results. Although the difference is significant according to a chi-squared test of the

error count data (χ^2 = (1, N = 1760) = 4.04, p = .044, V = .050), the association between the two variables is weak according to Cramér's V.

The average error rate as a function of the sound type differs by 5 percent, with earcons performing slightly better. A chi-squared test indicates the difference to be significant with χ^2 = (1, N = 1760) = 9.57, p = .002, V = .076). Although the difference is not substantial on average, it should be noted that the "bass" earcon, consisting of a short melody with a bass timbre, was always correctly identified in all trials. False positive errors, occurring when a user mistakenly indicates that the *key* item was present in the scene when it was not, were below 5 percent on average, and committed only by 7 out of 22 participants.

For the spatial loudspeaker playback, the test subjects were asked to state from which loudspeaker they thought the *key* sample was being played. As illustrated in figure 4.3 The angle mismatch between actual and perceived direction has a resolution of 30 degrees, defined by the loudspeaker spacing. This mismatch is compensated for by front-back reversals which map both actual and perceived direction to lateral angles between -90 and 90 degrees. The average absolute angle mismatch is about 30 degrees, the mismatch does not differ substantially between earcon and speech playback, and seems to be independent of the lateral angle of the *key* item.

4.4.2 Task 2 – Comparing two lists, which contains more key items?

The distribution of *key* samples was randomised, with set A containing more instances of the *key* sample in 50 percent of the cases and set B containing more instances in the remaining 50 percent of the cases. Test subjects were not aware of the distribution. Therefore, guessing would result in a 66.7 percent error rate.

The error rates decrease with increasing onset delay between sound samples. A chi-squared contingency table test of error count data reveals the effect to be significant with χ^2 = (2, N = 2640) = 301.88, p < .001, V = .338). Pairwise comparisons with Holm-Bonferroni correction show a significant decrease of the error rate for an increase of the onset delay from 100 ms to 200 ms and from 200 ms to 400 ms (see table 4.3).

	ISOI [ms]			**Difference [%]**					**Playback**		**Type**	
	100	200	400	33.3 (3vs.4)	50 (2vs.3)	66.7 (3vs.5)	100 (3vs.6)	133.3 (3vs.7)	HP	LS	E	S
Trials	800	800	800	528	528	528	528	528	1320	1320	1320	1320
Errors	433	238	103	245	215	131	103	71	385	389	412	362
Errors [%]	49	27	12	48	41	25	20	13	29	29	31	27
p	<.001	<.001		.037	<.001	.045	.03		.898		.036	
V	.228	.194		.074	.169	.064	.082		.003		.042	

Table 4.3: Task 2: Error count table with results from Chi-squared tests indicating statistically significant differences between all adjacent columns of the independent variables, except for "playback type", i.e., diotic headphone and spatial loudspeaker playback. V denotes Cramér's V. Difference in percent are rounded to whole numbers.

	Number of *key* objects				
AR scene 1	3	2	3	3	3
AR scene 2	4	3	5	6	7
Relative Distance [%]	33.3%	50%	66.7%	100%	133.3%

Table 4.4: Task 2: Objects per scene and relative difference in *key* object counts between the scenes.

A similar relationship holds for the difference between the error rates and the relative difference between the number of *key* objects in both AR scenes. Table 4.4 lists how the difference percentages are achieved. The error rates for different scene compositions are significantly different (χ^2 = (4, N = 2640) = 216.94, p < .001, V = .287). Again, pairwise comparisons with Holm-Bonferroni correction reveal a significant decrease of the error rate for each increase of the relative difference between the number of *key* items (see table 4.3).

The average error rates under both playback conditions, i.e., diotic headphones and multichannel loudspeaker playback, are equal. A chi-squared test of error counts does not indicate a significant difference (χ^2 = (1, N = 2640) = 0.02, p = .898, V =.003). The average error rates as a function of the sound type differ by 4 percent, with speech performing slightly bet-

ter. The difference is not substantial, although a chi-squared test indicates the difference to be significant with χ^2 = (1, N = 2640) = 4.39, p = .036, V =.042).

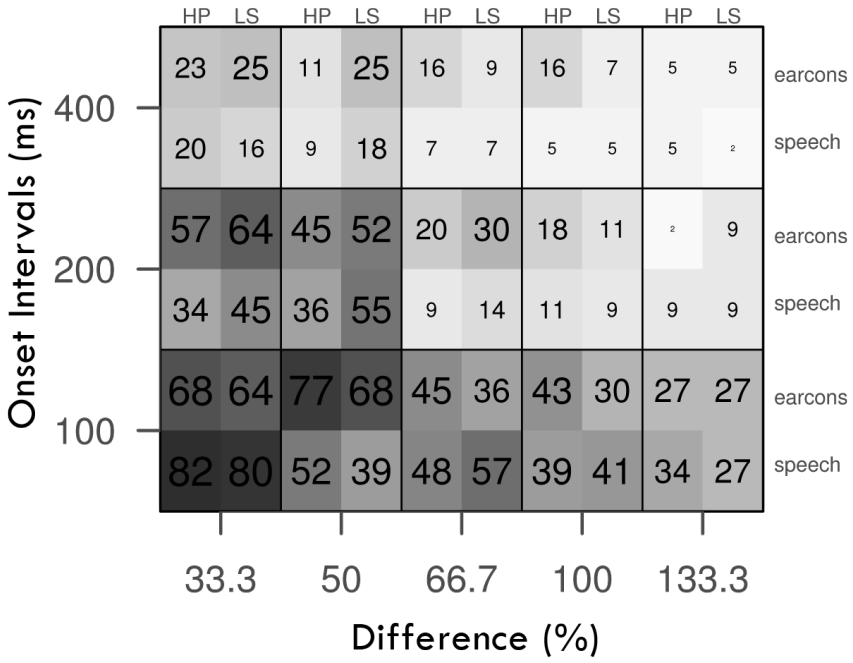


Figure 4.6: Task 2: Error rates (in percent) as a function of onset delays, playback type, sound type, and difference between number of target items.

For further insight into the data, the error rate was estimated using the regression tree analysis. It confirmed that the onset delay and the relative difference are the main determinants that affect user performance: the effect of other variables, including sound type and playback condition, was pruned out. For small relative differences (33.3 percent and 50 percent) and an onset delay of 100 ms the average error rate is 66.7 percent, which corresponds with random chance. Larger relative difference or onset delay improves performance substantially. If the onset delay and relative difference are at least 200 ms and 66.7 percent, respectively, the average error rate drops below 10 percent. Figure 4.6 gives an overview of error rates for all conditions. Note the low error rates in the upper right corner compared to the very high error rates in the lower left corner.

4.5 Discussion

4.5.1 Task 1

The error rates decrease significantly with each increase in the onset delay between consecutive sound samples. This finding implies that detection rates are relatively high even with a dense temporal presentation of the samples. With a minimum onset delay of 200 ms, the average error rates for determining the presence of the *key* sound sample drop below 10 percent. This finding is in line with hypothesis **H1**. Increasing the onset delay reduces the temporal overlap between consecutive samples and thus the total number of concurrently presented samples. This leads to improved user performance, but increases the overall playback duration.

No substantial differences were found in terms of the user performance between spatial loudspeaker and diotic headphone playback. Although earlier work has shown spatial separation to enhance the perception of concurrent sound sources [153], it did not have a substantial effect on the user performance for determining the presence of the *key* sample. A possible explanation could be that, unlike in the classical "cocktail party" situation, where the listener attends to sound coming from a certain direction, the direction of the *key* sample was randomised in every trial. To perform the task, the participant had to attend to sound originating from all directions, which may cancel out the advantage of spatial separation. The finding that detectability rates were not affected by spatial separation may be important for practical applications, where spatial separation of sound samples can be difficult or costly to implement.

As user performance did not differ substantially between synthesised speech and earcon playback, **H2** could not be confirmed. For the speech synthesis, words with similar tonal characteristics were chosen and the samples were generated using the same voice and the same parameters while the earcons offered more distinctive tonal features. The deliberately chosen low distinctiveness between the speech samples may have cancelled out potential advantages of using synthetic speech. Further improvement of the detectability rates for speech playback may be achieved by varying vocal characteristics, including gender, pitch, vocal tract size, accent, or speaking

style of the speech samples.

The "bass" earcon was always correctly identified. This can be explained by the fact that the energy of the bass earcon is concentrated at low frequencies, whereas for other earcons the energy was spread across the spectrum. The false positive error rate was less than 5 percent. This indicates that test subjects did not mistake the *key* sample with other samples present in the scene. However, when in doubt, test subjects seemed more likely to answer "not present" than "present".

H3 could be confirmed: The sound type and lateral angle of the *key* sample have no substantial effect on the localisation performance of test subjects. Once the *key* object had been identified, its location would be implicitly known, with an average absolute angle mismatch of about 30 degrees.

4.5.2 Task 2

The effect of the test conditions on user performance in task 2 is similar to the findings of task 1. **H1** could be confirmed: Each increase of the onset delay between consecutive samples significantly improved the performance for comparing the number of *key* samples in two lists. Increasing the onset delay reduces the temporal overlap of samples, and thus facilitates the recognition of individual samples. This improved the listeners' ability to understand the total number of *key* samples in one list scene relative to another. The effect may be due to psychoacoustic or spectral masking effects (see section 2.1.4), however, the experimental setup was not designed to study the causes in more detail.

Another factor affecting the user performance is the relative difference between the number of *key* objects in each list. Each increase of the relative difference led to a significant improvement in the participants' performance. This result is expected in **H4**: large numerosity differences are hypothesised to be easier to detect than small differences.

No difference in terms of the average performance between spatially separated sound samples played back via the spatial loudspeaker system and diotic headphones playback was found. Participants' ability to obtain an overview of the number of *key* items present in the lists was not affected

by the spatial quality of the sound samples. As in task 1, the directions of the *key* and distractor samples were randomised and unknown to the user a priori, which nullified the advantage afforded by spatial separation. Investigation of whether or not pre-existing knowledge of the direction from which a sample is played leads to an increase in user performance is an interesting topic for further research.

User performance did not differ substantially between earcons and speech playback. In analogy to the results of task 1, this indicates that the user can obtain an overview of the number of <key> samples in a set from synthesised speech samples with a similar accuracy as that achieved with earcons. As in task 1, varying the 37 vocal characteristics of the speech samples or the earcon design might further improve performance.

The regression tree analysis verified that the onset delay and relative difference are indeed the main determinants of the participants' performance in the tasks. The regression tree analysis identified a threshold near 60 percent relative difference (i.e., 3 versus 5 *key* samples): Differences above this threshold (66.7-133.3 percent, i.e., 3 *key* samples in one list versus 5, 6 or 7 *key* samples in the other) were detected substantially better than smaller differences (33.3-50 percent, i.e., 2 versus 3, and 3 versus 4 *key* samples), the effect being significant. The results indicate that neither sound type nor playback condition had a substantial effect on user performance.

4.6 Conclusion

This chapter presented results from a listening test concerning item detection and overview information in lists. Each item in the list presented to the test subjects was encoded as a short sound sample. Subjects were asked to determine the presence of a certain sample in a list (task 1), or to compare the number of occurrences of a certain sample in two lists (task 2). The parameters tested are derived from some of the design factors relevant for constituting auditory displays: information encoding, efficient and effective playback. The effect of these design parameters on user performance was studied by comparing speech and earcon sounds, a range of stimulus onset asynchronies (SOAs), and diotic headphone and spatial loudspeaker playback

in terms of user error rates.

These superordinate research questions were addressed in this chapter: **RQ 1.1**: How can spatial sound be utilised in an eyes-free interface? **RQ 1.2**: What are the advantages and disadvantages of using spatial sound compared to stereophonic or monophonic sound? and **RQ 3**: What is a good way to help users to obtain an overview of available items and options?

It was shown that audio is a viable candidate for item detection in lists and for comparing the numerosity of items in different lists. When the representation parameters are chosen suitably the error rates can drop to below 10 percent. Although onset delays of more than 200 ms significantly increase the performance, ISOI of 200 ms qualify as an acceptable trade-off between the length of the playback and the user performance.

Somewhat surprisingly the spatial information did not substantially increase or decrease performance. As a consequence, in practical applications, diotic playback may be sufficient, at least if explicit spatial information is not necessary. When spatial information is available the location of the sound source can be pinpointed moderately accurately, without affecting the user performance in the other tasks.

Another interesting observation is that the speech and earcons (except the "bass" earcon) performed comparably. This suggests that, all else being equal, synthesized speech is a good way to encode information, because it is easy to produce and the semantic is understandable without learning.

The study was designed to derive practical guidelines for auditory interface designers. For the benefit of practical relevance and to derive design guidelines, the experimental setup chosen for this study purposefully differs from psychoacoustic studies which usually focus on a specific cognitive process. While spectral or energetic masking (see section 2.1.4) or auditory memory effects (see section 2.1.3) may be highly influential factors, this experimental setup was not designed to investigate the impact of these factors on the results. The psychoacoustic mechanisms that lead to the finding, however, are an interesting topic for further study. Although this study was designed with a clear focus, the results obtained can be applied to a range of other scenarios, such as auditory representations of menus or files, or for optimising the echo metaphor proposed in chapter 3.

Chapter V

Simulator Sickness in 3D Audio Interfaces

5.1 Introduction

The design prototypes presented in chapters 7 and 8 rely on the unhindered perception of the simulated sound spaces and their movements or, on the movements of objects within them. Three-dimensional auditory interfaces can create a convincing illusion of "real" spatial sound and even whole simulated sound scenes including objects moving or being moved by a user. When the user and the auditory simulation move independently, the discrepancy between simulated motion and physical stillness can have unwanted side effects, such as the phenomenon of *simulator sickness* where a conflict between the motion perceived by the visual or auditory system and the vestibular system's sense of movement lead to symptoms of nausea, disorientation, and sickness. Simulator sickness has been comprehensively researched for visual experiences but there has been little research on auditorily induced simulator sickness. To identify its potential as a confounding variable for further experiments, this chapter addresses this subject in detail.

Vection, is thought to be one of the major candidates for causing simulator sickness [117, 118]. This illusionary perception of self-motion can occur in many real life situations – usually when an observer is still, but is exposed to a moving visual pattern, for example when watching a moving train through the windows of a stationary train, or seeing a film in the front rows of a movie theatre [118]. Vection can be facilitated by the perception of spatial presence in a virtual environment and the feeling of immersion [107, 119] (Chapter 6 addresses the potential for three-dimensional sound to induce the sense of presence and immersion in detail.). Studies concerning vection often assume a link between the vection measured and the potential for the device

or environment to cause sickness.

In this chapter the issue of simulator sickness induced by auditory stimuli is explored. The study presented addresses **RQ 1:** – What are the advantages and disadvantages of using sound? – by exploring the influence of movement patterns within a 3D sound space on the perceived pleasantness of the experience. Also, the influence of the perceived pleasantness/simulator sickness on the cognitive load generated by a simple task was studied. In section 2.4 a short introduction to the syndrome of simulator sickness is given. Johnson [211] is recommended for a more thorough introduction to the subject and an overview of research on visually induced simulator sickness.

This chapter is structured as follows: the related research section sketches out a summary of the work on vection, followed by an overview of research in the field of auditorily induced simulator sickness. The design rationale for the study is then described in detail. A summary of the findings is given and the results of the study are discussed. The chapter concludes with suggestions for future research on this aspect of auditory interfaces.

5.2 Related Work

While vection is a thoroughly researched phenomenon for visual stimuli, very few researchers have investigated vection in audio-only simulators or interfaces. Vection occurs for all motion directions and along all motion axes. In a typical visually induced vection experiment, participants are seated inside a optokinetic drum and are asked to report on their perception of self motion and their discomfort level. Most participants quickly perceive vection in the direction opposite to the drum's true rotation. Depending on the type of simulator used, over 60 percent of participants can experience motion sickness-like symptoms [212, 213, 214].

As found by Brandt et al. and Pausch et al. [215, 216], visual stimuli covering a large part of the field of view will usually induce stronger circular vection with shorter onset latencies. Stimulation of the entire field of view will result in strongest vection. Auditory vection has been less thoroughly researched. Although initial research was conducted almost 90 years ago [217], only recently has there been an increased interest in the phe-

nomenon. See [218, 219, 220] and a review by Väljamäe [221] for a comprehensive overview of auditorily-induced vection research.

A summary of the key findings: Lackner [222] found that a rotating sound field generated by either an array of six loudspeakers or dichotic stimulation can induce illusionary self-rotation, especially if the subject has their eyes shut or covered. See also Riecke et al. [223] for vection induced by moving sounds. If the subject has their eyes open and a stable visual field is given, vection does not occur. This and other research [224] suggests that visual cues dominate auditory cues in determining apparent body orientation and sensory localization.

Al'tman et al. [225] found that faster sound source movement was associated with an increase in the illusion of head rotation. In their study the subject was seated on a rotating platform and had their eyes closed. The subject's head was fixed in an immobile position, while an impulse series was played to them binaurally via headphones. The moving sound image affected the subject's postural reactions and created an illusion of head rotation. When there were changes in the sound source movement, vection effects, such as the perceived rotation speed, were particularly strong.

Larsson et al. [218] found that in a rotating sound field, sound sources associated with immovable objects (such as church bells) are more likely to induce vection than both moving (e.g. cars) and artificial sound sources. In addition they found that a realistically rendered environment may increase the perception of self-motion. The playback of multiple sound sources also induces significantly more vection responses than playing only a single sound source.

A rotating sound field or moving sounds, higher rotation speeds, changes in speed, sounds representing immovable sound sources, and a realistic sound scene are all factors intensifying the perception of vection. A stable visual field on the other hand decreases or impedes the perception of vection. Although several studies have shown that vection can be evoked by auditory stimuli, it is important to keep in mind that vection is only one possible cause for simulator sickness.

The user study reported in this chapter does not aim to reproduce the findings summarized above. The primary intention was to investigate the

general effects, including effects similar to simulator sickness, on a listener who is exposed to a binaural listening experience. In particular the effects on a listener of predictable and unpredictable movement patterns of sound sources in a 3D audio space were investigated.

5.3 User Study

5.3.1 Design Rationale

The user study was designed to induce motion sickness or sensations of discomfort in participants through playback of binaural recordings of movements between several competing sound sources. Having a mobile user in mind, these are some of the applications this research is relevant for:

- Mobile sound spaces, that move relative to a user, as used in navigation support systems [179, 180, 181]
- Binaural media consumption such as listening to binaural recordings of concerts or audio plays
- Spatial mobile conferencing with attendants located in a spatial, navigatable sound space [6].
- Spatial auditory interfaces that support navigation between and interaction with different sound items [14, 167, 7].

The user study was designed with the intention of delivering results with a practical relevance. Therefore, participants were neither blindfolded nor immobilized as in almost all settings, and most definitely in mobile usage settings, users will have visual stimuli, unless they are visually impaired, and it is unlikely that they would be unable to freely determine their body positions. If under these unrestricted, "realistic", conditions effects of simulator sickness arise, given the research summarised above, closing the eyes is likely to intensify, but not drastically change, the symptoms found by the study. The following conditions are compared:

Condition 1 (left-right): Predictable and consistent left-right audio movements of objects within a sound space which may occur while navigating, or interacting with a spatial audio interface.[1]

Condition 2 (random): Random audio movements of objects within a sound space which may occur during media consumption or live feeds from other users. These random movements are characterised by rapid changes of acceleration when approaching or withdrawing from a sound source and partial three-dimensional rotations.

Condition 3 (control): No audio movements; the control condition.

To study the effects of movement patterns on the perceived workload, participants were asked to identify random, nonsensical numbers in a text read to them (see section 5.3.4 for a detailed description). This task was designed to create a cognitive workload similar to the workload created by the challenges of orientation or navigating while focusing on a primary task. Based on the related research, the main hypotheses for this study were:

H1: Participants feel more discomfort when listening to random, unpredictable audio movements. Given the findings by Lackner [222] and Al'tman et al. [225], random and unpredictable movements would intensify the perception of vection and hence the perceived discomfort that they induce. Predictable movements, on the other hand, would allow the subject to form a mental model, which may potentially hinder the illusion of selfmotion [221] and therefore the experience of simulator sickness, assuming that vection is one of the main causes for simulator sickness.

H2: The distraction generated by random audio movements affects the cognitive load and decreases task performance. As unpredictable movements impede the formation of a mental model of a scene and its movement pattern, it was hypothesized that in the condition with random movements more cognitive load is generated by the continuous necessity to orientate oneself within the scene. Therefore the subjects may not be able to focus their attention fully on the task, which may have a negative effect on task performance.

[1] The recording setup is described in more detail in section 5.3.3.

5.3.2 Participants

82 participants, which were recruited within the Nokia community and several sport clubs, volunteered for the user study. The average age of the participants was 33 years, with the youngest being 15 years and the oldest being 54 years; 49 were male, 33 female. All participants were native Finnish speakers. The study was a between subjects design with participants randomly allocated to the three conditions:

- left-right: N = 28
- random: N = 25
- control: N = 27[2]

Three participants reported minor hearing problems but were not excluded from the study due to the negligibility of their impairments.

5.3.3 Audio Material

Twenty minutes of binaurally recorded sound was used for the study. The recording was produced by an experimenter wearing an Augmented Reality Audio (ARA) headset depicted in figure 5.1. The ARA consists of binaural microphones, an amplifier/mixer, and linear headphones and is designed to be "hear through", i.e. they allow users to hear audio cues superimposed over the audio from the surrounding real world [9]. The ARA headset was chosen for the recording instead of a manikin as it allowed the experimenter to move freely during the recording, which was especially important for recordings of random, 6-DOF movements. Binaural recordings with the ARA headset were preferred over binaural synthesis to ensure the reproduction of authentic head and body movements.

The task required participants to concentrate and stay focused on only one of the sound sources. Due to the inability of participants to exercise control over the movements of the environment, the perception that they themselves were motionless while the sound sourced moved around them was created. During the recording the experimenter sat on a swivel chair and

[2] Due to scheduling restrictions and cancellations an even distribution of participants between the groups could not be achieved.

Figure 5.1: ARA headset used in the study. Left: In-ear headphones equipped with microphones. Middle: Headset fitted in ear. Right: ARA mixer (Picture taken from [9]).

was surrounded by five Genelec 6020A bi-amplified active loudspeakers fixed at face level. The recording was made in a soundproof studio with room acoustics. As can be seen in 5.2, the loudspeakers were set up in a circular layout with a diameter of approximately 3 meters. The sound field created by the loudspeakers consisted of the following:

- Music, easy listening (Loudspeaker 1)
- Male speaker reading the text for the task (Finnish) (LS 2)
- Street noise, including cars passing by (LS 3)
- Podcast (Finnish), male and female speakers (LS 4)
- Environmental noise, birds, river (LS 5)

For the recording of condition 1 (left-right), the sound source the participants were asked to concentrate on the sound that was being played from loudspeaker number 2 shown in figure 5.2. For the recording the experimenter moved her head from left to right through an angle of 80 degrees over approximately 0.8 seconds as illustrated in figure 5.3. Preliminary testing indicated that having the target sound source positioned behind the participant is perceived as less natural and hence more annoying than a positioning in which the participant is facing the sound source.

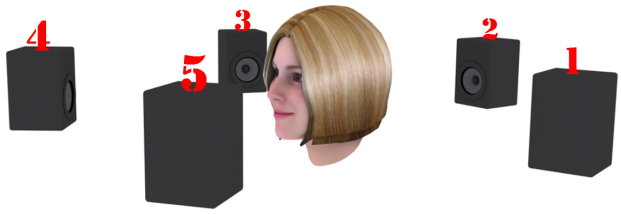


Figure 5.2: Setup used for binaural recordings with the experimenter surrounded by five loudspeakers.

For condition 2 (random) the experimenter moved her head using random, unpredictable movements. These movements included approaching or withdrawing from a loudspeaker, partial rotations about her x- ,y-, and z-axis, and rapid changes of acceleration during movements. For the control condition the experimenter faced the target loudspeaker and did not move at all.

Figure 5.3: Illustration of head orientation movements made for the recording of the left-right condition.

5.3.4 Task

In the study subjects wore headphones in a sound proof booth and were asked to listen to the recordings previously made in the different conditions.

In all conditions participants were asked to concentrate on one of the sound sources, a male voice talking about dogs and horses. The script was read by a professional male speaker and consisted of adaptations of the Wikipedia Finland entries on dogs[3] and horses[4]. 33 numbers between 1 and 120 were inserted in the text, care was taken to ensure that the numbers did not make sense in the given context. Participants were asked to identify the nonsensical numbers and write them down in the order in which they were presented during the test.

5.3.5 Procedure

Before their trial, participants were familiarized with the sound proof listening booths and were instructed on how to put on and adjust the Sennheiser HD580 headphones. After these instructions they were asked to fill the Simulator Sickness Questionnaire (SSQ) developed by Kennedy et al. [113]. They were then given an oral and written explanation of the task and encouraged to talk to the experimenter if they had questions or needed further clarification. After the trial participants were asked to fill out the SSQ a second time, followed by a questionnaire on their perception of various aspects of the study. After completing the questionnaire, participants were debriefed, compensated with two cinema tickets, and dismissed. The following dependent measures taken were:

- The pleasantness of the experience (including simulator sickness)
- The subjective perception of the sound space
- The task related error rate (interpreted as cognitive load)

The data from the various dependent measures were mostly analysed using a one-way analysis of variance (ANOVA) with a fixed confidence level (p-value = .05). Subjective ratings in the post-study questionnaire were scored on a seven-point Likert scale (From 1 – *I totally agree* to 7 – *I totally disagree.* Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the data set.

[3] `http://fi.wikipedia.org/wiki/Koira`

[4] `http://fi.wikipedia.org/wiki/Hevonen`

5.4 Results

The results presented in the following paragraphs are deduced from the data gathered through the SSQ, the error rate of the task, and the second questionnaire.

5.4.1 Simulator Sickness Questionnaire (SSQ)

The SSQ was used as a measure in this study. The symptoms used and their weightings are given in table 5.1. Sub-scales of the SSQ are: nausea, oculomotor, and disorientation. The oculomotor sub-score was removed from the questionnaire to adapt the results to measuring simulator sickness using a purely auditory stimulus.

Participants reported the extent to which they experienced each of the symptoms shown in table 5.1 as one of *None*, *Slight*, *Moderate*, and *Severe* before and after the trial. These were scored as 0, 1, 2, or 3. The sub-scales of the SSQ were computed by summing the scores for the component items of each sub-scale. Weighted scale scores, as specified by [113], were individually computed for each column by multiplying the *Nausea* scale score by 9.54 and the disorientation sub-scale by 13.92. The total SSQ score was obtained by adding *Nausea* and *Disorientation* values and multiplying by 3.74. As can be seen in figure 5.4, 51.9 percent of all participants had a score of zero or below for the SSQ *Total*, indicating that they did not show any symptoms of simulator sickness. However, approx. 48 percent of all participants showed slight to moderate symptoms. Tables 5.2 and 5.3 show pre-exposure scores, post-exposure scores and differences between the post- and pre-scores.

Symptom	**Severity**			
General discomfort	None	Slight	Moderate	Severe
Fatigue	None	Slight	Moderate	Severe
Headache	None	Slight	Moderate	Severe
Eye strain	None	Slight	Moderate	Severe
Difficulty focusing	None	Slight	Moderate	Severe
Increased salivation	None	Slight	Moderate	Severe
Sweating	None	Slight	Moderate	Severe
Nausea	None	Slight	Moderate	Severe
Difficulty concentrating	None	Slight	Moderate	Severe
"Fullness of the head"	None	Slight	Moderate	Severe
Blurred vision	None	Slight	Moderate	Severe
Dizzy (eyes open)	None	Slight	Moderate	Severe
Dizzy (eyes closed)	None	Slight	Moderate	Severe
Vertigo (Giddiness)	None	Slight	Moderate	Severe
Stomach awareness	None	Slight	Moderate	Severe
Burping	None	Slight	Moderate	Severe

Table 5.1: The Simulator Sickness Questionnaire (SSQ) used as a measure in this study, including the symptoms and their weightings.

Nausea

A *paired t-test* showed a significant difference ($t(26) = -4.24$, $p < .001$) between the pre ($M = 1.26$, $SD = 1.56$) and the post ($M = 2.56$, $SD = 2.12$) exposure scores for *Nausea* in the left-right condition. It also showed a significant difference ($t(26) = -2.76$, $p = .01$) between pre ($M = .081$, $SD = .8$) and post ($M = 1.69$, $SD = 1.95$) exposure scores for *nausea* in the control condition. However, the results from an *analysis of variance* on the scores for each condition shown in table 5.2 did not indicate significant differences in perceived *Nausea* between the conditions.

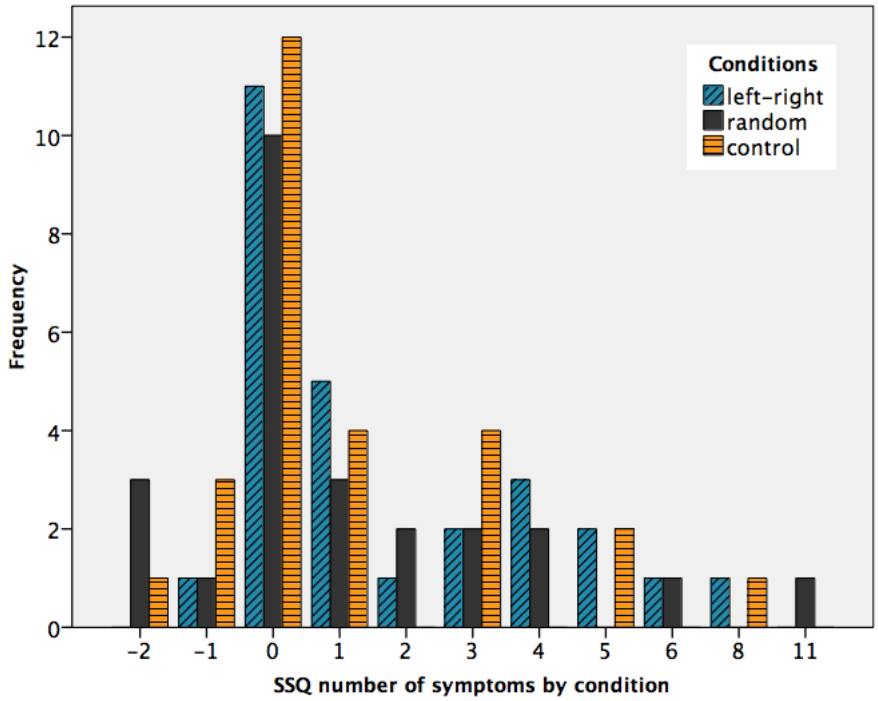


Figure 5.4: Frequencies for SSQ *Total* (unweighted) for all participants (N = 79) per condition.

Disorientation

A *paired t-test* showed a significant difference (t(24) = -2.28, p = .032) between pre (M = .36, SD = .57) and post (M = 1.12, SD = 1.81) exposure scores for *Disorientation* in the random condition. However, the results from an *analysis of variance* on the mean scores for each condition shown in table 5.3 did not indicate significant differences in perceived *Disorientation* between the conditions.

SSQ Total

For the left-right condition the SSQ *Total* is 6.65 (weighted), for the random condition it is 4.79 (weighted) and for the control condition 4.02 (weighted) (see figure 5.5). An *analysis of variance* did not show a significant difference (F(2,77) = .58, p = .56) for SSQ *Total* between the three conditions.

Condition	Nausea Pre (M)	SD	Nausea Post (M)	SD	Nausea Post-Pre (M, weighted)	SD
Left-Right	1.26	1.56	2.56	2.12	12.37	15.16
Random	1.08	1.23	1.6	1.58	4.96	13.53
Control	.81	.8	1.69	1.95	8.44	14.92

Table 5.2: Mean pre- and post exposure SSQ scores for *Nausea* over all three conditions.

Condition	Disorient. Pre (M)	SD	Disorient. Post (M)	SD	Disorient. Post-Pre (M, weighted)	SD
Left-Right	1.04	1.67	1.52	2.65	6.7	18.66
Random	.36	.57	1.12	1.81	10.58	23.18
Control	.42	.81	.69	1.34	3.61	13.15

Table 5.3: Pre- and post exposure SSQ scores for *Disorientation* over all three conditions.

5.4.2 Pleasantness

In the post-study questionnaire participants were asked to agree or disagree with statements about the general pleasantness of the experience. This was done on a Likert scale of 1 – I totally agree to 7 – I totally disagree. The included statements were:

- The task was pleasant.
- The listening experience was good.
- I could have continued to listen to this for a longer period of time.
- I would have liked to quit the test before the end.
- The sound volume was just right.

As Likert scales deliver an interval-level measurement the data were analysed using a one-way analysis of variance with a fixed confidence level (p-value = .05). Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the

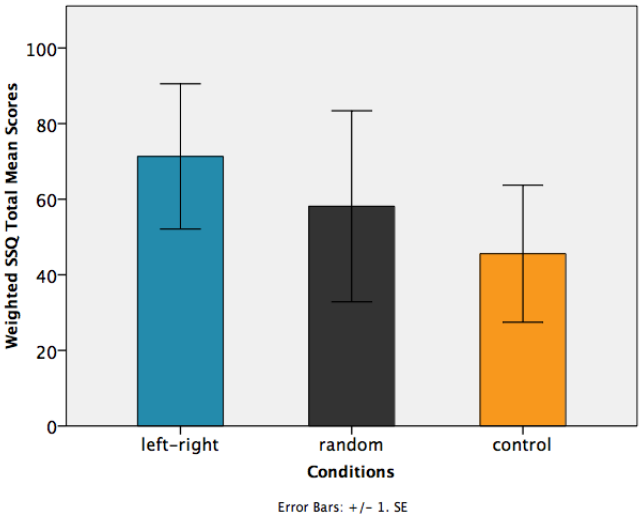


Figure 5.5: Mean scores for SSQ *Total* over all conditions. Higher values indicate stronger perceived simulator sickness.

data set. The number of participants, mean scores, and standard deviations are summarized in table 5.4.

Participants in the control group were on average indifferent about the pleasantness. As depicted in figure 5.6 participants from the left-right and the random group found the task to be significantly more unpleasant ($F(2,77) = 5.39$, $p < .01$ compared to the control group, which was confirmed by a post-hoc Bonferroni test (with $p = .02$ for left-right and $p = .01$ for random).

In response to the statement "The experience was nice/good.", participants in the left-right group found the listening experience significantly worse ($F(2,77) = 3.23$, $p < .05$, confirmed by a post-hoc Bonferroni test with $p < .05$) than participants in the control group (see figure 5.7).

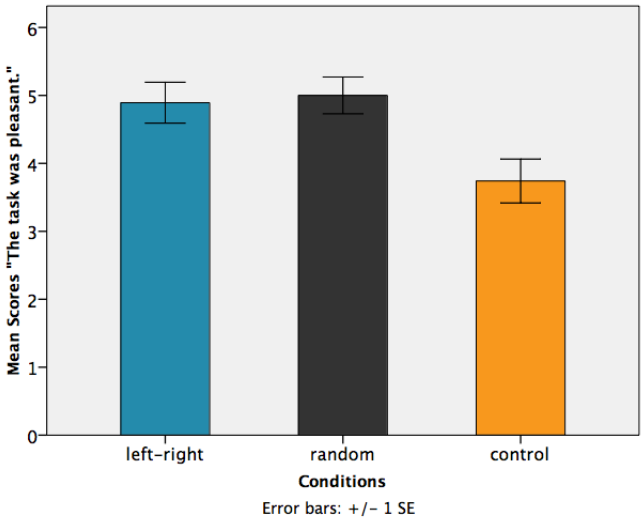


Figure 5.6: Mean scores for answers to the statement "The task was pleasant". Lower values indicate agreement.

In response to "I could have continued to listen to this for a longer period of time." over all conditions, participants felt they would not want to listen to the sound space for a longer period of time, although participants from the left-right group had significantly higher scores ($F(2,77) = 4.32$, $p < .05$, confirmed by a post-hoc Bonferroni test with $p < .05$) (see figure 5.8) than participants from the control group.

5.4.3 Perception of the Sound Space

Participants were asked whether they perceive the sound space to be chaotic. As can be seen in figure 5.9, participants in the control group ($N = 27$, $M = 3.37$, $SD = 1.85$) found the sound space to be significantly less chaotic ($F(2,77) = 6.67$, $p < .01$, confirmed by a post-hoc Bonferroni test with $p = .03$ for random and .002 for left-right) than participants in the left-right ($N = 27$, $Mean = 2.07$, $SD = 1.12$) and random ($N = 25$, $Mean = 2.36$, $SD = .95$) groups.

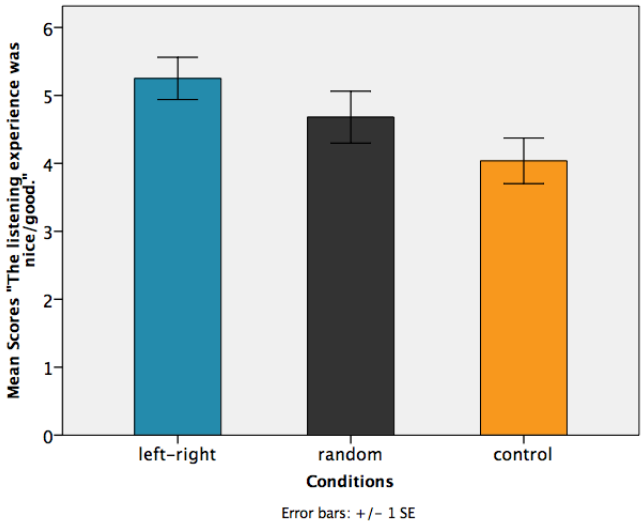


Figure 5.7: Mean scores for answers to the statement "The experience was nice/good". Lower values indicate agreement.

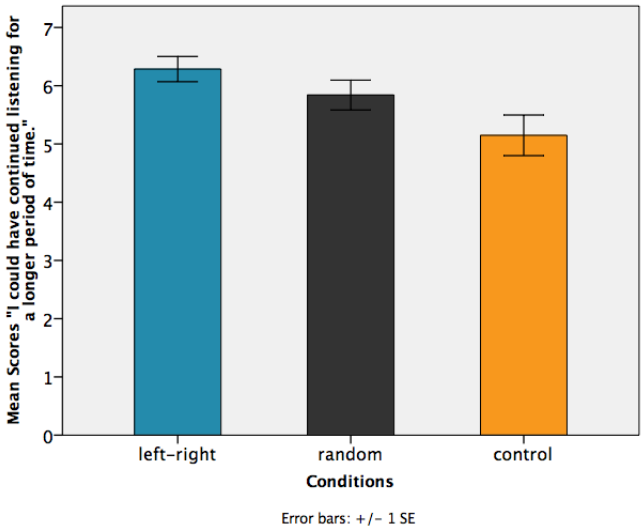


Figure 5.8: Mean scores for the item "I could have continued to listen to this for a longer period of time". Lower values indicate agreement.

Condition	**N**	**Mean Score**	**SD**
"The task was pleasant."			
Left-Right	28	4.89	1.60
Random	25	5.00	1.36
Control	27	3.74	1.50
"The listening experience was nice/good."			
Left-Right	28	5.25	1.65
Random	25	4.68	1.91
Control	27	4.04	1.74
"I could have continued to listen to this for a longer period of time."			
Left-Right	28	6.29	1.15
Random	25	5.85	1.28
Control	27	5.15	1.82
"I would have liked to quit the test before the end."			
Left-Right	28	4.32	1.87
Random	25	4.80	2.10
Control	27	4.85	1.90
"The sound volume was just right."			
Left-Right	28	2.25	1.18
Random	25	1.72	.74
Control	27	2.22	1.25

Table 5.4: Results from the post-study questionnaire on single items concerning the *pleasantness* of the experience.

5.4.4 Cognitive Load

To measure the cognitive load of participants during the trial the results from the listening task were evaluated. 33 nonsensical numbers were randomly inserted into the text and subjects were asked to write down the numbers. A comparison of the amount of numbers recorded across conditions showed no significant differences, in fact the results are almost identical. For control (N = 25) the mean of detected nonsensical numbers is 31 (SD = 2.4), for left-right (N = 27) the mean is 30.7 (SD = 3.1) and for random (N = 25) the mean is 30.9 (SD = 4).

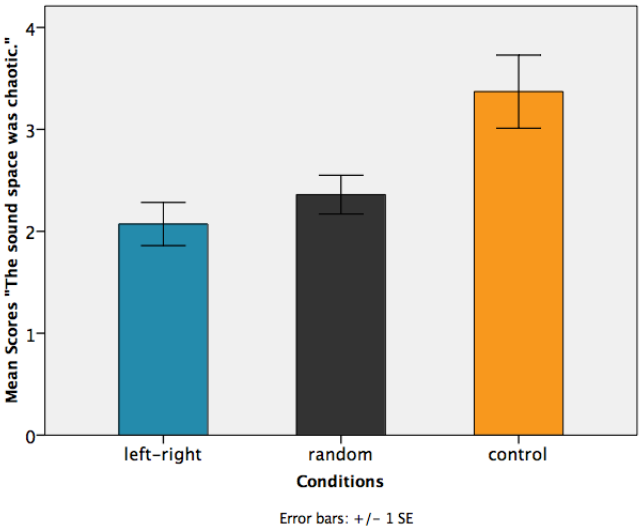


Figure 5.9: Mean scores for the perceived disorder of the sound scene. Lower scores indicate a higher perceived disorder.

In the questionnaire participants were asked whether they had found it difficult to concentrate on the task. The results shown in table 5.5 mirror the results from the evaluation of the task – participants were rather undecided, but showed a tendency in the random and control conditions towards having had more difficulty concentrating on the task.

Overall, participants did not have difficulties completing the task. This appraisal is supported by low mean scores (moderately strong agreement) for the statement "The task was easy" and for "The task was boring" throughout all three conditions. No significant differences were found between the conditions, but there is a tendency towards participants finding the left-right condition more difficult (see table 5.5).

5.4.5 Gender Differences

Evidence for different perceptions of both the task and the sound space between men and women were found in this study. Women (N = 32, M = 3.84, SD = 1.74) found it significantly more difficult (t(78) = -1.93, p = .05) to concentrate on the task than men did (N = 48, M = 4.56, SD = 1.56). Both women and men did not want to listen to the sound space for a longer period

Condition	**N**	**Mean Score**	**SD**
"It was difficult to concentrate on the task."			
Left-Right	28	4.04	1.67
Random	25	4.40	1.50
Control	27	4.41	1.87
"The task was easy."			
Left-Right	27	3.22	1.55
Random	25	3.08	1.58
Control	27	2.44	1.50
"The task was boring."			
Left-Right	28	3.71	2.12
Random	25	3.08	1.55
Control	27	2.85	1.38

Table 5.5: Results from the post-study questionnaire on how difficult participants rated the task.

of time. However, women (M = 6.31, SD = .93) disagreed significantly more strongly (t(78) = -3.1, $p < .05$) with the statement "I could have continued to listen to this for a longer period of time.".

Both men and women found the volume level to be very good, nevertheless men (M = 1.85, SD = .80) perceived it to be significantly better (t(78) = -2.04, $p < .05$) than the women (M = 2.41, SD = 1.39). Figure 5.10 illustrates the differences in weighted mean values for *Disorientation* pre-study, post-study, and in total. It also shows the differences between men and women for the SSQ *Total* score, including *nausea*. Results from an independent t-test show a significant difference (t(79) = 1.31, $p < .05$) between men (M = 4.54, SD = 14.87) and women (M = 10, SD = 22.7) in perceived *Disorientation* (weighted).

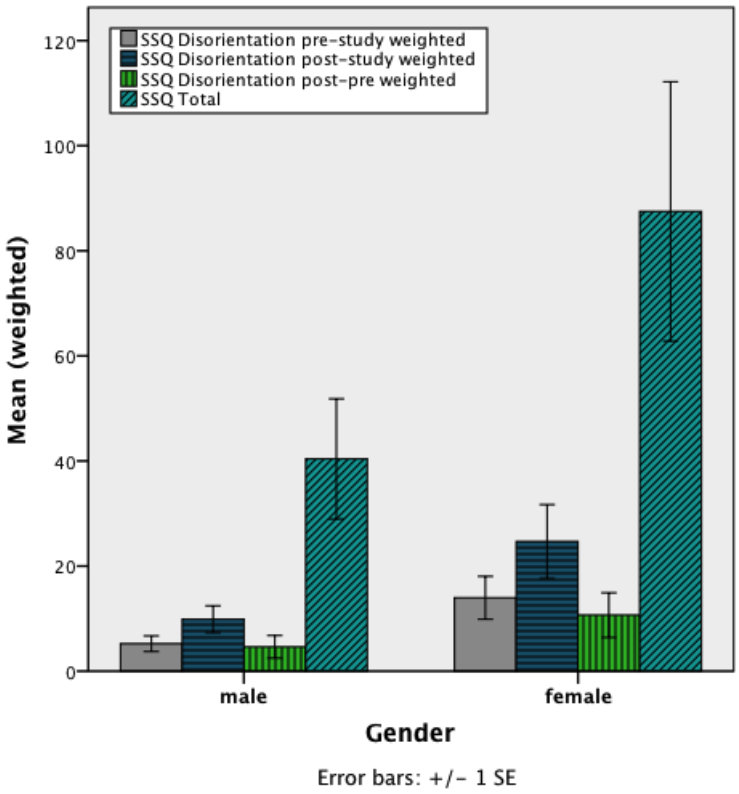


Figure 5.10: Weighted mean scores of men and women for *Disorientation* measured before and after the study, including the overall score for *Disorientation* (post-pre score) and the cumulative SSQ *Total* including *nausea* scores. Lower values indicate weaker symptoms.

5.5 Discussion

This study has demonstrated that different movement patterns of spatial sound sources do indeed affect the perceived pleasantness of the listening experience. It was found that predictable left to right movements make the listening experience less pleasant and generate stronger irritations compared to random movements or no movements at all.

The first hypothesis (H1) that random, unpredictable spatial sound movements would make the experience more unpleasant was not confirmed. A possible explanation may be that left-right movements in sequence was the most unnatural pattern. Humans use these movements to orientate themselves when crossing a street or when turning their heads towards a sound source, but this is not done repeatedly over a period of 20 minutes.

Participants may have also found it particularly annoying and/or boring to listen to left-right movements for a rather long period, compared to the random patterns, which in their diversity may have felt more natural and may have offered more challenge and hence, a more positive distraction.

In the control condition with no movements, the sound space was perceived to be least chaotic. This finding is as anticipated as we assume that movements, especially random and unpredictable movements, cause a delay in forming a correct mental model of the sound space. Also, such randomness impedes the anticipation of movements leading to the perception of a more chaotic sound space.

Furthermore, the results do not support the second hypothesis (H2) assuming a difference between the conditions in terms of distraction generated by the sound space. Unpredictable movements do not have an effect on the ability to concentrate on one sound source that is discernibly different from the effect caused by predictable movements or no movements at all. Generally, participants found the task to be rather easy and made fewer errors than expected. Further investigation is needed to fully understand whether or not there is a difference in cognitive load between the conditions and the extent of its significance for future developments in this field.

Results from the SSQ showed significant differences between scores from before and after the trial throughout all conditions, especially for the subscore *Nausea*. Although an analysis of variance did not indicate significant differences in the perceived *Nausea* or *Disorientation* between the conditions, the SSQ *Total* showed a trend towards a higher total score for the left-right condition. Although this finding is supported by results from the post-study questionnaire, given the high standard deviations (indicating a low consistency for the gathered data) it is difficult to make viable assumptions on the basis of this.

It was observed that some participants reacted strongly and showed severe symptoms of simulator sickness whereas others did not show any symptoms at all and indeed actually felt better after the study than they did before. Additionally, differences were found in the perception of the sound space between men and women. Women found it more difficult to concentrate on the task and they had a stronger dislike for the sound space. This perception was not confirmed by their task performance as there were no significant differences found between men and women. Although earlier findings by Kennedy et al. [226] and Biocca [227] indicate that women are more susceptible to simulator sickness, the data suggests that the differences in *Total* and *Disorientation* scores are likely due to a previously existing difference.

5.6 Conclusion and Future Work

In this chapter **RQ 1:** – What are the advantages and disadvantages of using sound? was addressed. It has been shown that predictable left to right movements lead to a perceived unpleasantness that is significantly higher than the unpleasantness experienced for unpredictable movements or no movements at all. Approximately 48 percent of all participants experienced mild to moderate symptoms of simulator sickness, with a trend towards stronger symptoms for the left to right movements. Although the consistency of the data gathered using the SSQ was very low, data from the post-study questionnaire and our observations indicate that spatial audio and especially regular movement patterns repeated for a longer period of time will have a negative impact on the perception of the sound space.

There are several directions for future research. In general, raising the consistency of the SSQ dataset should be aimed for. Varying the difficulty of the task used in the study as well as its realism should be considered in further investigations. Although some criteria of a mobile usage scenario (open eyes, no fixed posture) were acknowledged, others were neglected. Using a realistic mobile setting, for example an outdoors navigation task, could yield interesting insights. Furthermore, it would be interesting to investigate the user perception of spatial augmented reality audio applications, where a real sound environment is superimposed with a virtual auditory environment.

Considering that the study was designed with the intention of evoking symptoms of simulator sickness, the current data suggests that even under the extreme conditions created for the study the perceived unpleasantness that study participants experienced did not exceed an amount that would have lead to cessation of the trials. The unpredictable movements of sound sources in the sound space do not seem to reduce the listening experience to a critical degree, and there is no evidence of a negative effect on cognitive load for simple tasks. Therefore, designers and developers can be rather optimistic about the use of spatial audio in mobile applications, such as navigation support systems, spatial auditory interfaces or entertainment applications.

Chapter VI

Presence, Social Presence, Immersion

This chapter addresses **RQ 1.2**: What are the advantages and disadvantages of using spatial sound compared to stereophonic or monophonic sound?

It has already been shown that three-dimensional sound has the potential to add benefits to a variety of dialogue based applications. While listeners to mono or stereo sound often perceive the sound source to be positioned inside their heads, binaurally recorded or synthesized spatial sound externalises the position of the sound source. Spatial audio allows for substantial improvements in distinguishing and therefore following individual speakers in multitalker environments [68, 228].

As mobile hand-held communication devices recently started supporting three-dimensional audio by implementing spatial sound libraries such as OpenAL[1], the use of spatialised audio for speech based social, collaborative, and gaming applications is likely to increase. For these types of applications, immersion and the perception of social presence are considered core components of the overall experience.

Biocca & Harms [93] define social presence as a sense of being with another in a mediated environment. Both Biocca & Harms and Lombard & Ditton [92] point out that both the sense of accessibility to the emotional and intentional state of the other and, emotional interdependence contribute to the perception of social presence. Although there is a larger and very diverse range of different dialogue based applications, the underlying sensory immersion, i.e. immersion induced by the quality of the sound itself, not the content or type of application, is comparable.

[1] `http://connect.creativelabs.com/openal/`

It stands to reason that acoustic realism, including spatiality, might be a highly influential factor for the perception of (tele-)presence and social presence, but only few researchers have investigated this assumption.

The study presented in this chapter addresses RQ 1.2 by exploring potential difference between spatial and non-spatial sound in terms of the perception of presence and the understanding of the emotional state of several other speakers. An understanding of whether or not the playback quality of human speech hinders or facilitates the experience of presence is sought. Additionally, the issue of whether playback quality impacts the perception of being socially present in the (mediated) environment is investigated. For a more detailed introduction to the concepts of presence and social presence the reader is referred to section 2.3 of this thesis. Additionally, [107, 229, 230] are recommended for an overview on the topic. For the purposes of the presented research, the definition of social presence by Biocca & Harms is adopted

> [...] the sense of being with another in a mediated environment [...] the moment-to-moment awareness of co-presence of a mediated body and the sense of accessibility of the other being's psychological, emotional, and intentional states [93]

as well as the definition of presence proposed by Heeter [90] and Barfield et al. [231], i.e. the sense and feeling of "being there" in a mediated scene or virtual environment.

6.1 Related Work

Hendrix & Barfield compared the effects of spatialised and non-spatialised non-speech sound in a virtual environment on the user's perception of presence. They found that the addition of spatialised sound significantly increases the reported level of presence but does not have an impact on the perceived realism of the virtual environment [96].

Freeman & Lessiter [232], however, could not verify increased presence ratings for multi-channel audio when presenting participants with a rally car video sequence with accompanying synchronised audio. They argue, though, that these findings may be due to the lack of perceivable advantage of the

multi-channel presentation over the stereo presentation. Nevertheless, they found that enhancing the bass content and sound pressure level increases presence ratings.

Västfjäll [233] presented participants with music comparing the effect of mono, stereo, and six-channel sound on the participants' emotional reactions and ratings of presence. Music was chosen to convey either strong positive or strong negative emotions. Västfjäll found that both stereo and six-loudspeaker conditions were significantly more effective than the mono condition in inducing emotional reactions. Ratings for presence were significantly higher in the six-loudspeaker condition compared to the two other conditions. He concludes that presence is linked to spatial sound reproduction and emotional reactions vary as a function of the "immersivity" of the sound field.

Despite the comprehensive amount of research on presence in mixed and audio only (virtual) environments, surprisingly, the effects of spatialised and non-spatialised human speech on the perception of presence and the understanding of the emotional state of speakers have not been actively studied. The study reported in this paper is designed to gain insights into whether similar effects as those found for musical and environmental sounds can be found for human speech.

6.2 User Study

The user study was primarily designed to investigate whether or not there are differences in the human perception of monophonic, stereophonic, and binaural sound regarding the way in which humans perceive verbally communicated emotions. Furthermore, it was of interest to identify the differences – and any typology therein – in feelings of presence, i.e. having the perception of sitting among people conversing, rather than of sitting in a listening booth.

6.2.1 Design Rationale

To emotionalize participants and create an experience that induced the perception of being somewhere else, a scenario was chosen that most participants

were familiar with either from literature, film, television, or from personal experience. With the help of a professional scriptwriter a typical "confession scene" was created, in which Paula, Heikki's long-term partner, confesses to having an affair with Esa; the confession is made in reaction to Heikki's proposal of marriage. The emotionality of the scene is amplified by Esa's presence at the table.

As all test subjects were native Finnish speakers, the script was written in Finnish. To support a smooth transition from the unfamiliarity of the laboratory environment into the atmosphere of the scene, the play started as a regular dinner invitation, with Paula and Heikki as hosts and Antti and Esa as their guests. During the first third of the play all characters are introduced. The atmosphere is friendly, relaxed and cheerful. The positive emotional climax is reached when Heikki proposes to Paula, immediately followed by the turning point of Paula's confession. The second third is dominated by Heikki's feelings of utter surprise, incredulity, and later anger as well as Paula's feelings of shame and guilt – colliding in a heated discussion. Emotions calm down during the last third of the play but remain unresolved. The play ends with Esa and Antti being asked to leave and their compliance with Heikki's request. The experiment compared three conditions:

- Condition 1: A monophonic recording
- Condition 2: A stereophonic recording
- Condition 3: A binaural recording

As repeated exposure to the content would have diminished the element of surprise and presumably the level of emotionality, a between-subjects experimental design was chosen. The study was designed to test the following hypotheses:

H1: The stereophonic and binaural conditions differ significantly in terms of perceived presence from the mono condition. The perceived presence is strongest in the binaural condition. It was assumed that feeling present in a "virtual" environment requires a sense of spatiality and the layout of this environment. Hence it was surmised that the binaural condition, offering more spatial information, would outweigh both other conditions in terms of perceived presence.

H2: The stereophonic and binaural conditions differ significantly in terms of the understanding of the emotions acted out in the play. The understanding and alignment is stronger in the stereo and binaural conditions than in the mono condition. Given Västfjäll's [233] findings, it was assumed that, similar to his findings for music reproduction, speech based audio may induce stronger emotional reactions in both stereo and binaural reproduction than in mono.

The questionnaire used to evaluate the listening experience comprised basic demographic questions, questions about the sound quality and the participants' emotional state, as well as several items taken from the questionnaire on Mediated Communication Experience (ComXQ) [234]. Furthermore, participants were asked to represent their perception of the scene through a sketch.

6.2.2 Audio Material and Recording Technique

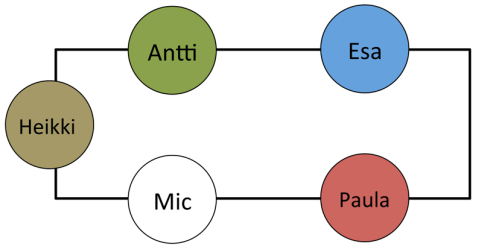


Figure 6.1: Illustration of the character's seating order.

To ensure maximum quality of the audio play, four professional actors from the Tampere Komediateatteri[2] performed the play. As can be seen in figure 6.1, the actors were seated at a table. Figures 6.1 and 6.2 show Heikki as the host at the head site of the table (far left), to his left Antti and Esa, to his right the (imaginary) listener – represented by a HEAD acoustics HMS II.3 manikin and several microphones – and Paula (far right).

Several props were used during the recording, e.g. a bottle of wine, wine glasses, plates, cutlery, and music ("easy listening", which faded out by the end of the first third). The play was recorded with multiple microphones in a recording studio fulfilling the requirements set in ITU-R BS.1116[3].

[2] `http://www.komediateatteri.fi/`

[3] Recommendations of the International Telecommunication Union: Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems: `http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-`

Figure 6.2: Actors during the recordings of the play.

The background noise level was minimal and reverberation times were typical of those found in a large living room. The audio capture was done on a computer located in the next room running Adobe Audition 3.0 in multitrack mode. The audio card used was RME Hammerfall DSP Multiface II. Presonus Firestudio was used as an additional ADAT A/D converter. The Presonus's internal microphone pre-amplifiers were used for the five main recordings. The mono recording was captured with a RØDE NT2-A microphone located in front of the manikin (see Figure 6.3). The microphone was set to use an omnidirectional polar pattern. It was located slightly below the manikin's ear level so as not to distort the binaural recording.

The stereo recording was done with an ORTF stereo capture configuration also with RØDE NT2-A microphones, which were set to a cardioid polar pattern. In ORTF capture the microphone capsules are located 17 centimeters from each other and spread at a 110 degrees angle. ORTF provides both volume difference (with signals arriving at cardiod microphones at different angles) and timing difference as the sound arrives at the separate microphones with different delays. The binaural recording was done with a HEAD acoustics HMS II.3 artificial head and torso simulator. A type 3.4 artificial ear according to ITU-T Rec. P.57[4] was used for recording.

1-199710-I!!PDF-E.pdf

[4] http://www.itu.int/rec/T-REC-P.57/en

Figure 6.3: Manikin used for the recording with a mono RØDE NT2-A in front of the mouth and a stereo RØDE NT2-A ORTF-pair just above the head pointing outwards.

During the experiment participants listened to the recordings in silent isolation booths [235] with Sennheiser HD-580 headphones. The volume level was set to be same for all recorded configurations.

6.2.3 Participants

82 participants volunteered for the experiment ranging in age from 15 to 54 years (M = 33 years), and were recruited within the community of a large company and several sport clubs. 49 participants were male, 33 female. All participants were native Finnish speakers. Participants were randomly allocated to the three conditions. Three participants reported having minor hearing problems. They were not excluded from the experiment due to the negligibility of their hearing problems. Two participants did not complete the experiment.

6.2.4 Procedure

Before their trial, participants were asked to sign a consent form and were briefed about the nature of the experiment. They were then familiarized with the listening booths and were instructed on how to put on and adjust the headphones. Subsequently they were asked to sit down in their assigned booth, relax, and focus on what they were about to hear. After the trial, participants were asked to fill out a questionnaire.

6.2.5 Experimental Design

A between-subjects design was used for this experiment. The 80 valid participants were randomly assigned to one of three groups. Participants in group one listened to the monophonic recording, group two to the stereophonic recording, and group three to the binaural recording. Group one comprised 25, group two 29, and group three 26 valid participants.[5] No explicit task was given to the participants other than to relax and focus on the conversation.

6.3 Results

A seven-point Likert scale has been used in the questionnaire (1 = I totally agree and 7 = I totally disagree). The questionnaire was in Finnish and all items discussed below are translations of the original items. As the data were normally distributed and Likert scales deliver an interval-level measurement the data were analysed using a one-way analysis of variance (ANOVA) with a fixed confidence level (p-value = .05) unless otherwise stated. Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the data set.

As one of the larger subsets of questions was designed to evaluate the listening experience in terms of perceived presence and emotional alignment to the content of the audio play, this subset of eighteen questions was examined for underlying dimensions reflecting these constructs. A Principal Component Analysis (PCA) over a subset of eighteen variables was run and three

[5] Due to scheduling restrictions and cancellations an even distribution of participants between the groups could not be achieved.

variables with commonalities $< .6$ could be excluded according to stepwise principles. Finally, the PCA was run again on fifteen variables (KMO = .79, Bartlett's Test of Sphericity $< .001$, Varimax Rotation). As there were fewer than thirty variables and commonalities after extraction greater than .6, all factors with Eigenvalues above 1 (Kaiser's criterion) were retained. The PCA result indicated four factors explaining 68.5 percent of the total variance. These factors were interpreted as:

1. Presence (five variables)
2. Emotional Understanding/Involvement (five variables)
3. Focus (two variables)
4. Authenticity (three variables)

A another PCA was run on a second subset of the dataset. From the initial 18 variables four variables were excluded using the same criteria as stated above. The PCA (KMO = .74, Bartlett's Test of Sphericity $< .001$, Varimax Rotation) results indicated that three factors explained 63.6 percent of the total variance. These factors were interpreted as:

1. Negative Emotions (six variables)
2. Positive Emotions (five variables)
3. Alertness (three variables)

Due to the small sample size and the rather exploratory nature of the items chosen for the experiment the sum score method was used to further analyse the data [236]. Summed factor scores preserve the variation in the original data, which is useful for the further analysis. Only items suggested by the PCA (with a loading of $< .6$) were used. Cross-loading items were assigned to the score they loaded higher on. All items on a factor were given equal weight, regardless of the loading value.

6.3.1 Presence

The *Presence* construct (Cronbach's alpha: .79) combines the following questions:

1. I felt like the participants in the conversation surrounded me.

2. I felt like I could reach out and touch the participants in the conversation.
3. I felt I was face-to-face with the participants in the conversation.
4. I would have liked to actively participate in the conversation.
5. I felt more like a participant than an observer of the conversation.

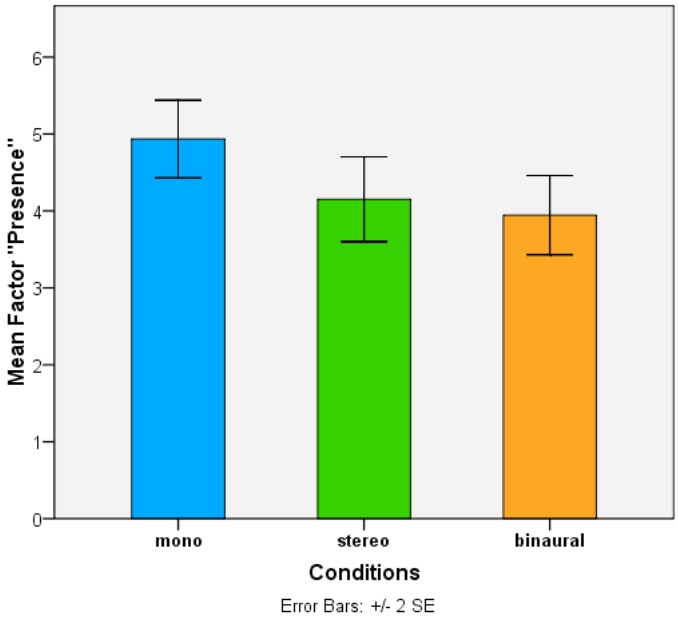
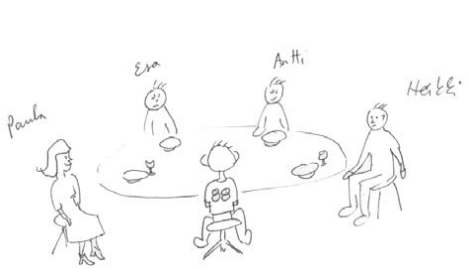


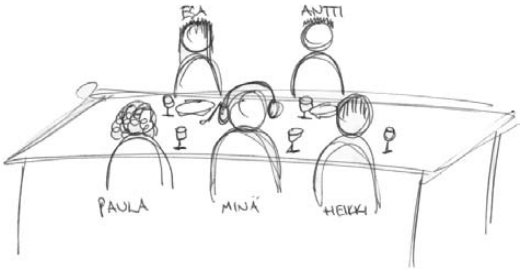
Figure 6.4: Mean values for the factor *Presence* over all three conditions. Small values represent a stronger sense of *Presence*.

A significant difference ($F(2,77) = 3.74$, $p = .028$) between the conditions was found. A post-hoc Bonferroni test (with $p = .034$) showed these differences to be between the mono ($N = 25$, Mean = 4.94, SD = 1.26) and the binaural ($N = 26$, Mean = 3.95, SD = 1.31) condition. No significant difference between the stereo condition ($N = 29$, Mean = 4.15, SD = 1.49) and the mono ($p = .116$) or binaural ($p = .98$) condition was found.

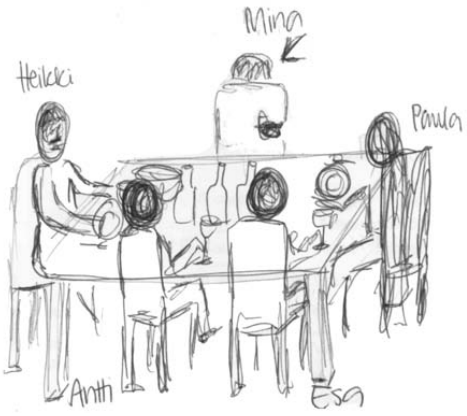
As illustrated by figure 6.4, participants in the binaural group agreed significantly more strongly with the statements listed above and hence had a stronger sense of *Presence* than participants from the mono group. Besides being asked to respond to questions, participants were prompted to sketch the situation they just listened to. Figures 6.5 and 6.6 show a representative



(a) Participant No. 88 is sitting among the characters.



(b) The participant draws himself as wearing wearing headphones.



(c) Participant is sitting with his/her back towards the table; Indication of a front-back-confusion.

Figure 6.5: Example of a drawing showing the participants sitting at the table among the characters of the play.

sample of the drawings.[6] Proxemic behaviour is one of the possible objective measures of co-presence identified by Biocca et al. [108]. Although the experimental setup did not allow for participants to move, the drawings are particularly useful for showing whether participants saw themselves as part of the group (as exemplified by figures 6.5a, 6.5b and 6.5c) or as observers (exemplified by figures 6.6a, 6.6b and 6.6c). These drawings are not only interesting in terms of the participants' sense of membership, but also, in some cases, as indicators of the experience of front-back-confusion[7]. Figures 6.5c and 6.6a point to perceived front-back-confusions.

[6] The often found label *minä* in the drawings translates to the English "me" or "I"

[7] Items that are located in front are mistaken for being located behind the listener.

(a) The participant draws herself as separate from the characters; Indication of a front-back-confusion.

(b) The participant is sitting in the audience of a theatre like room and is watching a play.

(c) The participant does not draw herself as part of the scene; The characters' emotions are depicted, showing Heikki enraged and Paula in tears.

Figure 6.6: Example of a drawings showing the participants separated from the group.

Some of the sketches even give an insight into how the emotional content has been perceived, as for example Figure 6.6c, in which Heikki, who had just learned that his partner Paula is having an affair with Esa, is shown angry and in an agitated pose, while Paula, who had just rejected Heikki's proposal and confessed her affair, is shown in tears.

The analysis of the drawings solely focused on the position of the participants and refrained from further interpretation. Participants drawing themselves as sitting at the same table and being part of the group were counted as "sitting among the speakers". Participants drawing themselves as

sitting at a distance or being otherwise physically separated were counted as "sitting or standing separate from the speakers". Drawings which were not depicting the scene or which were indecipherable were marked as "other". As participants were not specifically asked to draw the emotional state of characters – or where exactly they were seated – these data were not used in a comparative analysis.

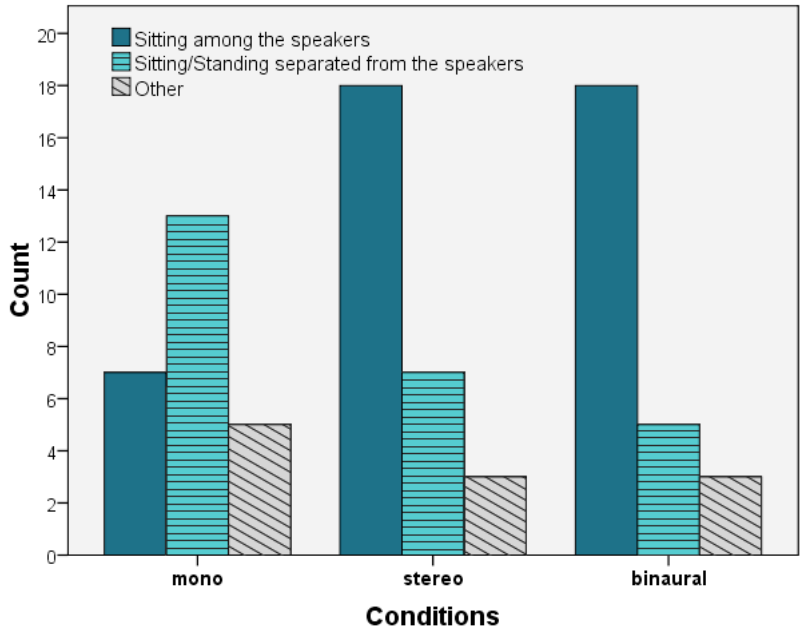


Figure 6.7: Counts over all three conditions for sketches depicting the participants as part of the group or as observers.

As mentioned above, the analysis of the sketches supports the results from the analysis of variance conducted on the factor *Presence*. A a χ^2-test was conducted to test the frequency distribution between the groups of participants drawing themselves as being seated among the speakers or sitting/standing at a distance. As illustrated in Figure 6.7 a relationship between the quality of the audio and the different depictions could be confirmed with $\chi^2 = (4, N = 79) = 10.7$, $p = .031$.

In conclusion, both the analysis of variance of the factor *Presence* and the interpretation and analysis of the sketches suggest that the sense of presence in a virtual scene or environment is significantly stronger when participants

listen to binaurally recorded sound compared to monophonic sound. The difference between binaural and stereophonic sound is not as distinct, but still verifiable.

6.3.2 Emotional Involvement / Understanding

The following questions were combined under the factor *Emotional Involvement/Understanding* (Cronbach's alpha: .8):

1. The mood of the participants affected me.
2. I identified myself with one or more of the participants.
3. I knew how the participants felt.
4. I was emotionally moved by the conversation.
5. I was immersed in the situation.

An analysis of variance showed a significant difference ($F(2,77) = 3.582$, $p = .033$) between the conditions. A post-hoc Bonferroni test ($p = .029$) showed the difference to be between the mono ($N = 25$, Mean = 3.48, $SD = 1.194$) and stereo ($N = 29$, Mean = 2.69, $SD = 1.009$) condition. As can be seen in figure 6.8, participants tended to agree more strongly with the statements listed above in the stereo condition than in the mono condition.

The binaural condition ($N = 26$, Mean = 2.97, $SD = 1.073$) tested as not significantly different from the mono ($p = .296$) or binaural ($p = 1$) condition. Participants in the stereo condition showed the highest emotional involvement/understanding. On average they reported having a good understanding of how the characters felt. They felt more immersed and more affected than the participants reporting on the other two conditions.

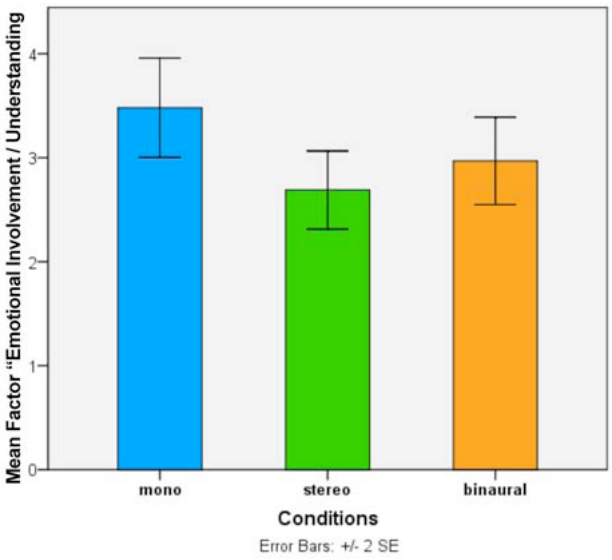


Figure 6.8: Mean scores for the factor *Emotional Alignment/Involvement* by condition. Lower scores indicate higher emotional alignment/involvement.

6.3.3 Focus

The factor *Focus* (Cronbach's alpha: .46) only comprises two questions, namely:

1. I was focused on the conversation and did not pay attention to the surroundings or the equipment.
2. The test environment did not diminish the listening experience.

An ANOVA revealed a significant difference between the conditions ($F(2,77) = 3.163$, $p = .048$). A post-hoc Bonferroni test ($p = .05$) showed the difference to be between the mono ($N = 25$, Mean = 3.52, SD = 1.617) and stereo ($N = 29$, Mean = 2.552, SD = 1.325) conditions. The binaural condition ($N = 26$, Mean = 3.289, SD = 1.537) was not significantly different from neither the mono ($p = 1$) nor the stereo ($p = .213$) condition.

As illustrated by Figure 6.9, generally participants found the test environment did not have a strong impact on their listening experience or distracted them from focusing on the audio play. However, the stereo condition had significantly lower means compared to the mono condition, indicating a positive impact of the stereo condition on the ability to focus on the play.

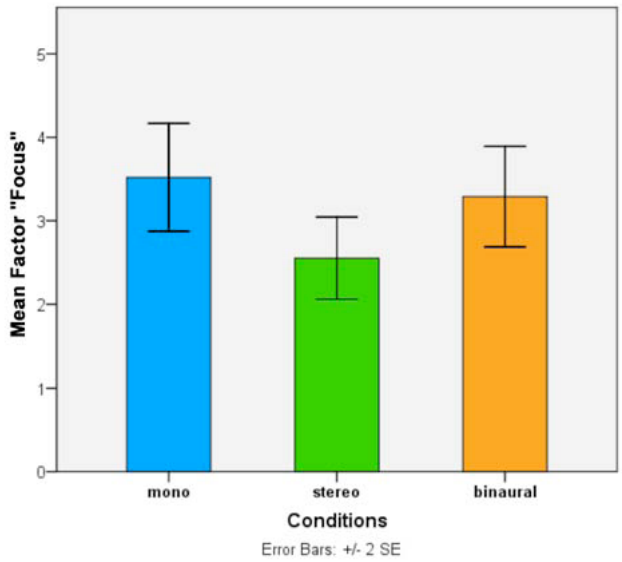


Figure 6.9: Mean scores for the factor *Focus* by condition. Lower scores indicate higher *Focus*.

6.3.4 Authenticity

Gathered in the factor *Authenticity* (Cronbach's alpha: .78) are the questions:

1. I enjoyed listening to the conversation.
2. The conversation was convincing.
3. The scene felt alive and vivid.

There were no significant differences between the conditions in respect to the perceived authenticity of the conversation. Although there is a trend towards a lower mean value in the stereo (N = 29, Mean = 2.598, SD = 1.448) and binaural (N = 26, Mean = 2.69, SD = 1.073) conditions compared to the mono (N = 25, Mean = 3.48, SD = 1.194) condition. In general participants believed the conversation to be quite authentic.

6.3.5 Emotions

In the post-study questionnaire participants were asked to agree or disagree (on a 7-point Likert scale) to eighteen questions about their emotional state. A PCA suggested a clustering into three constructs:

1. *Negative Emotions* ("I feel: unhappy/shocked/angry/frustrated/confused/-aggressive.") [Cronbach's alpha: .819]
2. *Positive Emotions* ("I feel: relaxed/happy/satisfied/cosy/chatty.") [Cronbach's alpha: .84]
3. *Alertness* ("I feel: active/lively/curious.") [Cronbach's alpha: .55]

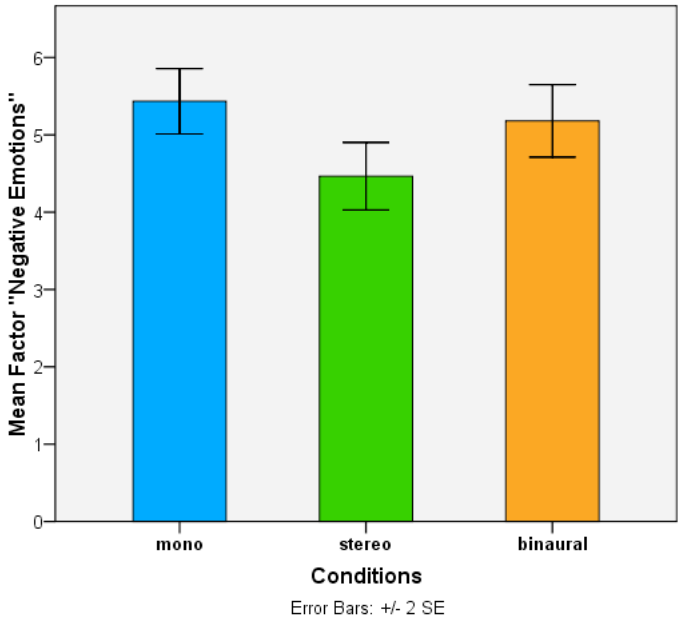


Figure 6.10: Mean scores for the factor *Negative Emotions* by condition. Lower scores indicate stronger negative Emotions.

An ANOVA showed a significant difference ($F(2,77) = 5.269$, $p = .007$) between the conditions for *Negative Emotions*. A post-hoc Bonferroni test ($p = .008$) indicated this difference to be between the mono ($N = 25$, $Mean = 5.433$, $SD = 1.057$) and the stereo condition ($N = 29$, $Mean = 4.47$, $SD = 1.174$). No significant difference between the binaural ($N = 26$, $Mean = 5.18$, $SD = 1.19$) and stereo ($p = .071$) or mono ($p = 1$) condition was found.

As illustrated in figure 6.10 participants generally tended to disagree when asked if they felt negative emotions. However, in the stereo condition participants disagreed less strongly when asked about their negative emotions.

There was no significant difference found between the conditions in respect to the factors *Positive Emotions* and *Alertness*. Participants generally felt indifferent (means of 3.5 for binaural and stereo and 3.8 for mono) when asked about their positive emotions. Participants tended to feel rather alert with mean values around 3.0 for the stereo and binaural conditions, and around 3.4 for the mono condition.

6.4 Discussion

The presented study was designed to provide insights into the effect of mono, stereo, and binaural sound on the perceived social presence as indicated by the perception of presence in a virtual scene and the understanding of the emotional state of speakers in the scene.

The results show that, as assumed in **H1**, there are significant differences between the conditions. Both methods used, the questionnaire and the visualization of the scene through drawing, showed a higher perceived presence in the binaural condition compared to the mono condition. The visualisation method proved to be an especially rich source not only for insights related to how participants perceived themselves in relation to the characters, but also for subtle signs of what they considered to be the predominant emotional content. Additionally, the drawings could be used as indicators for localization accuracy and front-back confusions.

Participants in the stereo condition showed the highest emotional understanding with a mean value of 2.69. On average they reported having a good understanding of how the characters felt. They felt more immersed and more affected than in the other two conditions. **H2** could only be partly accepted as a significant difference between the mono and stereo conditions was found, but no statistically relevant difference between the binaural and mono conditions. Given Västfjäll's findings, [233] no difference between stereo and binaural sound in terms of emotional involvement/understanding was assumed. Only a significant difference for the factor Negative Emotions between the mono and the stereo condition was found, not between the mono and the binaural condition. As the play was written to be emotive and as the emotions displayed in the play were predominately 'negative' emotions of anger,

shame, and fear, it is not surprising that only influence on negative emotions (as only these had been manipulated) was observed.

In contrast, it was unexpected that the binaural condition did not show a significant difference from the mono condition. It seems that the externalization of the sound source has no impact on the perception of other humans' emotions. There are several possible explanations for why the participants in the binaural condition showed weaker emotional alignments than hypothesized. Firstly, there might have been a greater number of" participants with lower accuracy in their localization ability ("bad localizers") in the binaural condition. This seems unlikely, though, as it would have affected their analysis on the factor *Presence*, which it did not. Secondly, there might have been unknown psychological effects involved. For example, as participants in the binaural condition had a stronger sense of presence, they may have dissociated themselves from the display of negative emotions as a defence mechanism and therefore may have shown lower emotional alignments. Thirdly, as the binaural listening experience through headphones was new to the participants, they may have been focusing on the medium to a greater extent than the message. However, as these are only speculations, further experimental examinations focussing on this factor are necessary.

The stereo condition was shown to have a positive impact on the ability to focus on the play. An explanation for this might be that as participants in the stereo condition showed stronger emotions and were more emotionally involved, they may have found it easier and more interesting to follow the conversation. In line with Hendrix & Barfield [96] no effect of the conditions on the perceived authenticity of the conversation was found.

During this experiment no tracking mechanisms were used that would have allowed participants to turn their heads and face individual speakers. Enabling certain forms of human social behaviour – like moving closer to speakers or facing them – might influence the perceived realism, presence, and the perception of displayed emotions, particularly in the binaural condition.

6.5 Conclusion

In this chapter **RQ 1.2**: What are the advantages and disadvantages of using spatial sound compared to stereophonic or monophonic sound? was addressed. Beyond the obvious benefits of spatial sound, i.e. its three-dimensionality, the observations in this chapter suggest that a designer should prefer spatial sound reproduction over stereo or mono reproduction if a strong sense of presence or social presence is desired. Furthermore, strong evidence was found to support the hypothesis that stereo provides a significant improvement – at least when there are multiple participants conversing – for enhancing the understanding of the emotional state of speakers. Unlike gaming environments, most mediated communication does not make use of stereo sound but is monophonic only. Adding a second channel could significantly improve the communication experience. As a consequence, the prototypes described in section 7.2 and chapter 8 make use of spatial sound to create an environment in which items can be selected, moved and manipulated. They use stereo sound for situations, in which a clear focus on a sound source is desired. For example in section 7.2 one person can be selected from a three-dimensional display of people conversing and then focused on by "expanding" their stream to a stereo stream and consequently muting all other streams.

Chapter VII

User Interaction with 3D audio Interfaces

While chapter 2 gave a broad introduction to auditory interfaces and related fields, other chapters have addressed research questions related to ways of improving distance perception (chapter 3), the efficiency of sound for item identification and overview techniques (chapter 4), the occurrence and impact of simulator sickness (chapter 5) and the influence of the recording and playback technique on the perceived amount of presence and social presence (chapter 6). In this chapter research is presented that addresses the following research questions:
RQ 2: What are viable non-visual multimodal interaction techniques?
RQ 2.1: What are the advantages and disadvantages of different tactile interaction techniques?
RQ 3: What is a good way to help users obtain an overview of available items and options?
RQ 5: How can the focus of attention be supported?

Special emphasis is put on investigating how users interact with auditory interfaces. While the WIMP paradigm still dominates visual interfaces and has only been slightly modified for the mobile domain, dominant input control and interaction strategies for auditory interfaces have yet to be established.

Mobile devices in particular provide new possibilities in term of tactile interaction due to the large range of embedded sensors and the devices' physical form factor. Shaer & Hornecker [237] give a thorough overview of work in the field of tangible interaction. Tactile interfaces such as vibration feedback have the advantage of not drawing the user's visual attention away from their main activity. However, they are mostly useful only for short notifications, not for communicating any complex messages. The user also needs to be

within reach of the device in order to register signals such as vibration.

Different types of gesture techniques or gesture taxonomies seem promising and have been proposed previously for hand-held devices see, for example [238]. However, there is still little knowledge about the acceptability and usage of some of these techniques, especially in the context of an auditory interface. Therefore, the first study summarized in section 7.1 investigates the design space of tangible interaction with a mobile auditory interface proposed and conceived by end-users. The results of this explorative study are discussed in terms of the scope of the gestures proposed, their tangible aspects and the user preferences. The results of this study deliver initial gesture recommendations that then influenced the follow-up studies described in section 7.2 and section 7.3.

The first empirical study described in section 7.2 compares gesture-based with key-based user interaction given an auditory interface. The study was designed to simulate an application that incorporates many of the metaphors[1] and interaction strategies established in existing human-computer interfaces. The notion of data being composited in a *file* represented by an *icon* that can be *selected* is such a metaphor, or the concepts of *minimizing/maximizing* to focus attention or the organisation of *files* into navigable *hierarchical data structures* are similarly metaphorical. The context of the first study was a mobile usage scenario assuming user interaction with a hand-held device, while the second study investigated user interaction with both an auditory and a visual head-down interface while driving a car. In this case the interaction is still tangible but is not based on gestures, as driver safety dictates minimizing tangible distraction (*"Eyes-on-the-road / Hands-on-the-wheel" paradigm*). Therefore, a prototypical steering wheel was built incorporating two mouse buttons and a scrolling wheel to provide a precise, user friendly and safe way of interaction.

This chapter has three main parts: The first section (7.1) is a description of an exploratory study on how users intuitively use a hand-held device to interact with an auditory interface. The results of the analysis are provided (section 7.1.5), followed by a summary of the findings and a discussion

[1] Please refer to section 2.6.4 in chapter 2 for an introduction and discussion of the use of metaphors in human-computer interfaces.

(section 7.1.7) of the implications of the study for interaction designers. In the second section (7.2) the results of a comparison between traditional keypad based interaction and a newer, tangible approach using the phone itself as a device for navigtion within a virtual spatial auditory environment are described and discussed. The findings from this study are presented in section 7.2.4 and are discussed in section 7.2.7. The third section (7.3) describes the second empirical study that focussed on a multimodal approach and compared user interaction with a visual and two auditory interfaces in a car. The results of this study can be found in section 7.3.5, followed by a discussion of the findings in section 7.3.6. The chapter concludes with a summary of the overall findings and recommendations for the design of tangible interaction schemes for mobile auditory interfaces.

7.1 Explorative Study on Tangible Interaction

7.1.1 Introduction

Recently, the manipulation of information on mobile devices has moved from keypad interaction to touch input. However, one of the drawbacks of this approach is that it requires a large part of the users' visual, cognitive, and motor attention that can be harmful, or not adapted to some specific mobile situations such as steering a car or walking in a busy urban environment. Speech recognition is a very straightforward approach for input control, but it is difficult to use in mobile situations due to the high signal-to-noise ratio caused by traffic, other conversations, and constantly changing environmental sound fields and levels.

Gesture interaction techniques that can exploit inbuilt sensors such as accelerometers, gyroscopes, and/or digital compasses can overcome the previously described issues by providing an elegant, eyes-free solution for user input [239]. The large range of potential gestures that can be executed in a mobile context and the different factors (DOF, user dexterity, mobile form factor, etc) do, however, need to be studied more thoroughly.

The exploratory study presented in this section aims to shed light on the type of gestures users will perform freely and intuitively when interacting with a spatial auditory interface. For this purpose participants in a qual-

itative user study were asked to perform several tasks. Their actions and comments were recorded, analysed, and are summarised in the results section (7.1.5). A discussion of the results can be found in subsection 7.1.7, followed by suggestions and guidelines for gesture design in section 7.1.8.

7.1.2 Related Work

Over the last few years we have seen the emergence of work on gesture techniques for spatial interaction design. One of the earliest mobile auditory interfaces presented was the Nomadic Radio introduced by Sawhney & Schmandt [14][2]. Gestures for mobile auditory interfaces have been explored in different projects and the validity of the approach has been shown in user studies like Pirhonen et al. [16]. In terms of techniques, Brewster et al. [167] presented one of the first spatialised audio systems combined with gestures. The Shoogle system [15] engaged the user to query information by shaking the device and delivered information using vibrotactile and sonified feedback. Similarly, Li et al. [177] presented a set of eyes-free gestures using a mobile keypad for a (non-spatial) auditory interface after surveying the most relevant tasks for end-users.

A recent study by Rico & Brewster [240] has observed the general social acceptability of some gesture techniques and demonstrated their impact for a real usage of some of the proposed metaphors. Similarly, Bhandari & Lim [241] asked participants to match specific gestures to common tasks realized on mobile devices. In contrast to the approach taken for this study, the gestures were predefined by the authors and not designed by the participants. Montero et al. [242] investigated how both users and bystanders feel about performing a particular interaction with a mobile device. They identify the following factors as the main contributors to the acceptance of a new interface: culture, time, interaction type, the user's position on the innovation-adoption curve, and the user's perception of the others' ability to understand the behaviour they are observing as being reasonable.

In summary, gesture techniques for mobile devices have been well devel-

[2] A detailed description of the projects mentioned in this section can be found in section 2.6.5.

oped and explored. Yet there are only few studies exploring gestures as an input technique for spatial auditory interfaces, and none of these studies actively involves end-users in a participatory design process for the development of these gestures.

7.1.3 Experimental Design

The motivation for this study was to investigate how users would interact with basic elements of a spatial auditory interface without restricting them by pre-defined gestures or limited system capabilities. Of special interest were the concepts and metaphors users would transfer from their everyday usage of computers and mobile phones to the entirely unknown domain of spatial auditory interfaces.

Ten participants were solicited; four male and six female. The mean age was 30, with participants spanning from 12 years to 49 years of age. Participants with a wide variety of professional backgrounds were chosen. All participants were familiar with desktop computers and were using them at least once a week for communication, accounting, or gaming. All participants owned a mobile phone but just one owned a smartphone.

Before the experiment participants were familiarized with synthesized spatial sound by playback of sound scenes consisting of single and multiple sound sources via headphones. Finally, the participants were given a featureless phone dummy made of wood to perform the gesture they would envisage for each of the different task (see figure 7.1). Users were allowed to perform any gestures with or on the device. The experiment was video recorded and, following the think-aloud protocol [243], participants were encouraged to verbalize their thoughts during each task. After the study participants were interviewed (unstructured), debriefed, and compensated with a cinema ticket.

7.1.4 Tasks

The tasks for the user study were based on a modified concept of the traditional WIMP-based desktop interface. As the WIMP paradigm is what participants were familiar with, they were asked to imagine spatialised sound

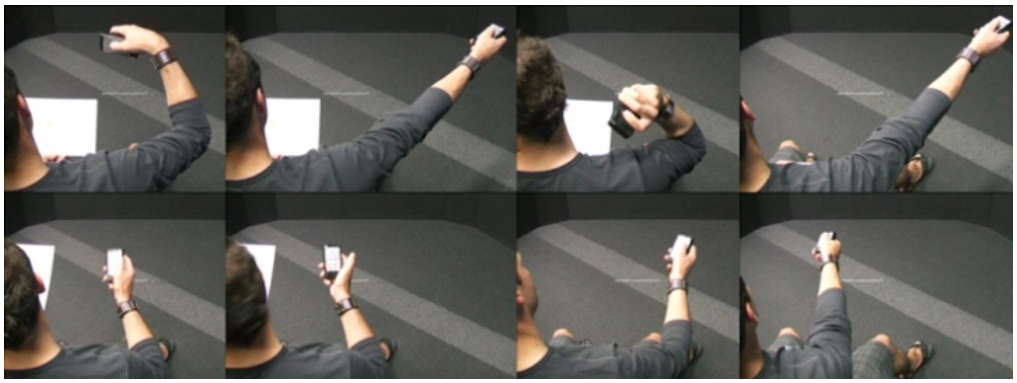


Figure 7.1: A participant performing a gesture for moving a sound source (task 14).

sources to represent applications, files, and folders, which would be selectable, manipulable, and structured hierarchically. Each of the tasks was explained symbolically and textually to the participants.

A total of 20 tasks were structured in three main categories:

Item Selection

Task 1: Select a single sound source.

Task 2: Select a sound source from a list.

Task 3: Skip through sound sources in a list.

Task 4: Deselect a selected sound source.

Task 5: Select several disjointed items from a list.

Task 6: Select several contiguous items from a list.

Task 7: Select all items of a list.

Attention Prioritization

Task 8: Change the distance of a sound source.

Task 9: Maximize/focus attention on one sound source.

Task 10: Undo maximization of one specific sound source.

Task 11: Minimize one specific sound source.

Task 12: Undo minimization of several sound sources.

Task 13: Minimize all sound sources.

Item Manipulation

Task 14: Move a single sound source.

Task 15: Lock a single sound source.
Task 16: Unlock a single sound source.
Task 17: Pause a single sound source.
Task 18: Re-activate a paused sound source.
Task 19: Delete a single sound source.
Task 20: Activate/open a single sound source.

7.1.5 Results

Participants used a total of 254 gestures. 98 of these were 3D movements with the device, 137 gestures were performed on the "touchscreen", and 19 were combinations of both 3D and 2D gestures. An overview of the most frequently used gestures is presented in table 7.1. Essential gestures are illustrated and described in detail along with user comments and some notes on the domains from which users transferred gestures to solve the tasks.

Touchscreen and embodied gestures

The pointing gesture **Point** was one of the most elemental 3D gestures used in order to select an item (see figure7.2, left). This gesture often preceded other gestures such as **TiltUp** (see figure 7.2, right) and **TiltDown**, **DoubleTouch**, **Arc**, etc.

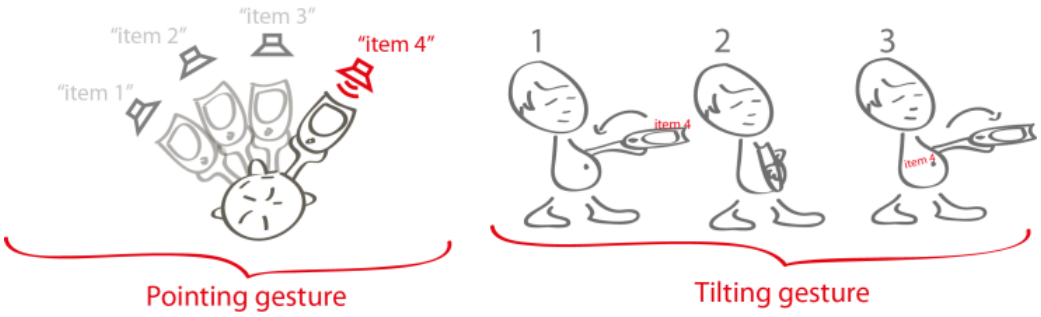


Figure 7.2: **Point** (left) and **TiltUp + Move** (right) gestures.

Some gestures were similar to gestures already implemented on the iPhone or Android Devices such as **ScrollUp**, **ScrollDown**, **Drag&Drop**. More original gestures proposed by the participants were:

- **Arc**: The device is moved from point A to B in an arc shaped curve
- **Flick**: A flicking hand movement – like throwing a Frisbee
- **PageFlip**: Rotating the device around its vertical axis
- **Shoot**: Pretend to shoot with the device

Disjointed item selections (like task 5) were usually realized by repeating the single-selection gestures from task 1 and 2. Most contiguous item selections (like task 6) were split into three different actions: Selection of the frist item + Browsing the list + Selection of the last item using 2D as well as 3D or combinations of gestures. With regard to reversible commands (like undo/redo): participants preferred repeating the original commands (cf. task 9 and 11) or reversed the gesture (e.g. **Arc towards user** and **Arc away from user**).

Combinations of 2D and 3D gestures

Combined gestures were mostly found in tasks involving 2D object selection (**Hold**) combined with spatial object manipulations such as moving a sound source using the 3D **PickUp&Drop** gesture. Figure 7.4 shows an example of a combination of a 3D gesture (**TiltUp**) followed by a 2D gesture (**ScrollDown**).

Figure 7.3: Selecting contiguous items with a combined gesture (task 6).

2D Gestures	**3D Gestures**
Task 1: Select a single sound source.	
Scanning the screen with a finger + Hold for selection of focussed item (1)	Point + Touch (9)
Task 2: Select a sound source from a list.	
Drag&Drop sidewards + Press (3) Drag&Drop sidwards + ScrollDown (3)	Point + TiltUp (2) Point + TiltDown (1) Point + TouchBelt[3] (1)
Task 3: Skip through sound sources in a list.	
Drag&Drop sidwards (4) ScrollUp + ScrollDown (2)	Shake left/right (3) PageFlip (1)
Task 4: Deselect a selected sound source.	
ScrollUp (3) Touch (2) Press (1)	Shake left/right (2) Flick (1)
Task 5: Select several disjointed items from a list.	
n×Drag&Drop sidewards + ScrollDown (5) n×Drag&Drop sidewards + Press (1)	n×(Point + TiltUp)(2) n×(Point + TiltDown)(1) n×(Point + TouchBelt)(1)
Task 6: Select several contiguous items from a list.	
Hold + n×(Drag&Drop + ScrollDown[4] (6) Touch + n×(Drag&Drop + Press&Hold (2)	PageFlip (1) TiltUp + PageFlip + TiltDown (1)

[3] See figure 7.2 for an illustration.

[4] See figure 7.3 for an illustration.

Task 7: Select all items of a list.

DoubleTouch (3)	Draw a circle [O] (3)
Press&Hold (1)	PointAll + TouchBelly (1)
Draw a circle [O](1)	
Drag&Drop+ScrollDown(1)	

Task 8: Change the distance of a sound source.

ScrollUp + ScrollDown(6)	TiltUp + TiltDown (4)

Task 9: Maximize/focus attention on one sound source.

DoubleTouch (4)	Device to landscape format (3)
Press&Hold (1)	Shake (1)
	Arc towards user (2)

Task 10: Undo maximization of one specific sound source.

DoubleTouch (4)	Shake (4)
Press&Hold (1)	Arc away from user (2)

Task 11: Minimize one specific sound source.

Multitouch zoom out (1)	Push away (2)
Draw a slash [\] (1)	Move downwards (2)
Flip away (1)	Cross out (1)

Task 12: Undo minimization of several sound sources.

DoubleTouch (2)	Pull close (2)
ScrollDown (1)	Move upwards (2)
	Shake (1)

Task 13: Minimize all sound sources.

Draw a circle [O] + ScrollDown (2)	Draw a circle [O] + TiltDown (2)
Draw a circle [O] + Draw a slash [\] (2)	PointUp + TiltDown (1)

Task 14: Move a single sound source.

Drag&Drop (3)	PickUp&Drop (8)

Task 15: Lock a single sound source.

Press&Hold (3)	Lock (turn a key in a door) (2)
Draw a circle [O] clockwise (2)	TiltUp (2)
	DoubleTiltUp (1)

Task 16: Unlock a single sound source.

Press&Hold (3)	Unlock (turn a key in a door) (2)
Draw a circle [O] counterclockwise (2)	TiltUp (2)
	DoubleTiltUp (1)

Task 17: Pause a single sound source.

Touch (4)	TiltDown (1)
DoubleTouch (2)	Move away (1)
Draw a slash [\] (2)	

Task 18: Re-activate a paused sound source.

Touch (4)	TiltUP (1)
DoubleTouch (2)	MoveCloser (1)
Draw a slash [/] (2)	

Task 19: Delete a single sound source.

Cross out [X] (2)	Cross out [X] (3)
Flip away (1)	Flick (2)
	Shoot (1)

Task 20: Activate/open a single sound source.

DoubleTouch (4)	Shake (2)
Touch (1)	TiltUp (2)
	TiltDown (1)

Table 7.1: Most frequently used 2D and 3D gestures by task. (Frequency of appearance is indicated by the number in parentheses.)

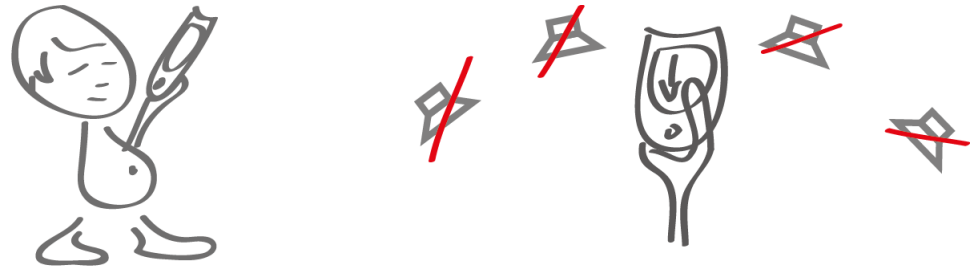


Figure 7.4: Combined gesture to mute/minimize all sound sources (task 13).

7.1.6 Gesture Associations

Participants were encouraged to think aloud during the experiment. Comments offered insights into the participants' associations and intentions. Some of the gestures based on previous experiences were:

- Windows 7 uses a **Shake** gesture to quickly minimize every open window except the one shaken: participant shook the device to focus attention on a sound source. Most operating systems have a ⊗ button to close an application: participants used the X-shape to delete a sound source.
- The iPhone and Android OS require the user slide a finger across the screen to scroll between different screens: participants used this to skip through sound sources in a list.
- Flipping pages in a book: participant used this to skip through sound sources in a list.

7.1.7 Discussion

Three main categories of gestures were used by participants: 2D gestures, 3D gestures and combinations of both. With rising task complexity a preference for 2D and 3D gesture combinations was noticed. Participants often used 2D gestures (e.g. **Hold**) to address an object and proceeded with a 3D gesture like **Tilt**, **Shake**, or **Move** to manipulate this object.

Half of the participants clearly preferred 2D gestures, while only few participants were using solely 3D gestures, but almost all participants combined gestures for certain tasks. From the think-aloud protocol it became obvious that feedback for successful (and unsuccessful) gestures is essential. Also, participants wished to have an overview of available commands executable on an object or container.

The gestures were mostly created through associations with interaction techniques known from other devices and tool. Experiences from the real physical world were mainly used in cases where the person did not have an association with a technical context (e.g. to "lock" a sound source). Although only one user owned a smartphone, several participants slid their fingers over the imaginary screen to skip through sound sources in a list or to scroll up and down (iPhone or Android OS touch screen gesture).

Other noticeable analogies were based on the users' current usage of traditional desktop GUIs (e.g. ⊗ icon to delete a sound source or the mouse shake to focus on a source). The **Touch**, **DoubleTouch**, and **Point** gestures show strong resemblance to using a remote control or a computer mouse (**DoubleClick** resembles the **DoubleTouch**).

Participants favoured reusing a set of basic gestures for different tasks (differentiated by a specific context) or 'inverse' gestures over having a wide range of unique gestures. An alternative strategy for keeping the gesture set small and simple is to use context menus, although this may slow down the interaction and may require more cognitive attention.

7.1.8 Conclusion

This section addresses two of the superordinate research questions: **RQ 2**: What are viable non-visual multimodal interaction techniques? and **RQ 2.1**: What are the advantages and disadvantages of different tactile interaction techniques? The result of this study is an overview of gestures users would perform with and/or on a handheld device to interact with basic elements of a spatial auditory interface. Users chose gestures based on pre-existing knowledge and the ability to translate experiences from other domains to the domain at hand. When the aim is to design an intuitive and user-centred

auditory interface, these are the main recommendations derived from the study:

- Use a small, context related gesture set
- Support gesture inversions for do-undo-commands
- Support gestural analogies from other domains
- Give clear and distinct feedback to actions
- Provide information about available commands
- Favour minimal gestures over expressive, elaborate gestures

It was also observed that participants generally preferred to use discreet gestures instead of very expressive gestures. Compared to the results gained in [241], where a preference for 2D gestures for public scenarios is advised, the same tendency was demonstrated in this study with the exception that participants found minimalistic 3D gestures to be acceptable.

7.2 Empirical Study 1: Gesture vs Key based Interaction

In this section an exploration into the usability of spatial sound and multimodal interaction techniques for a mobile phone is described using the example of a multiparty phone conferencing application. The study compares traditional keypad based interaction to that of a newer approach using the phone itself as a device to navigate within a virtual spatial auditory environment. Taking into account the findings of the study presented in the prior section 7.1, this study focusses on simple *yaw/panning* (see figure 7.7) and *pitch/tilting* (see 7.8) gestures performed with the phone.

This section is structured as follows: An introduction to the context of this study is given in section 7.2.1 followed by a short overview of related work and a summary of of recent research findings in section 7.2.2. The experimental setup and the description of the experimental procedure can be found in section 7.2.3. Section 7.2.4 reports the results, which are then discussed in section 7.2.7. Based on these results, some general conclusions are drawn in section 7.2.8.

7.2.1 Introduction

Due to limitations in phone hardware, until recently remote participants were 'plugged into' a phone conference using monophonic audio streams. Although there were multiple sound streams from the various participants, they were – and mostly still are – channelled through a single audio output, making it difficult for a listener to distinguish between speakers.

Previous research has shown that spatial sound cues can be used to distinguish between multiple sound sources, improve speech perception, and facilitate speaker identification [244, 245, 246]. With recent developments in the smartphone market and devices supporting spatial sound libraries like OpenAL[5], it is possible to harness the benefits of spatial audio to improve the quality of multiparty mobile calls.

[5] `http://connect.creativelabs.com/openal/`

However, before this can be realised, research is required on how to present these multiple streams of information, and on how to design appropriate interaction methods for such non visual tasks.

The original research interest in undertaking this study emerged from the desire to understand how motion tracking can be utilised to provide eyes-free interaction with complex user interfaces. In this follow-on study, the objective was to further explore the use and the efficiency of gesture-based interaction techniques for item selection, item manipulation, and sound stream monitoring in a multitasking environment. It was of special interest to further explore how well interaction paradigms for visual interfaces and point and click devices translate into a spatialised auditory domain for the mobile user.

7.2.2 Related Work

Several researchers have explored the uses of spatial audiovisual cues for stationary and mobile applications. For example, Crispien & Savidis [170, 247] designed an egocentric spatial interface for navigating in, and selecting from, a hierarchical menu structure[6].

Kobayashi & Schmandt [168] built an egocentric dynamic soundscape to create a browsing environment for audio recordings. Frauenberger & Stockman [169] positioned the user in the middle of a virtual room with a large horizontal dial in front of them. The menu items are presented on the edge of the dial facing the user while the rest of the dial disappears behind a wall. The user can turn the dial in either direction by using a gamepad controller. Sawhney & Schmandt [14] created one of the first mobile spatial audio interfaces – the so-called "Nomadic Radio", a spatial audio application based on a wearable computer. In the Nomadic Radio audio messages are positioned in a circle around the listener's head according to their time of arrival. User interaction was by means of voice commands and tactile input.

Walker & Brewster [248] developed single-user spatial audio applications for PDAs and mobile phones. Their work showed how spatial audio can be

[6] A detailed description of the projects mentioned in this section can be found in section 2.6.5.

used effectively for information display on a PDA to overcome the limitations of a small screen. They used spatial audio to convey time remaining on a file download using an audio progress bar. Brewster et al. [167] created a mobile system based on Audio Windows by Cohen and Ludwig [162]. They used spatialised auditory icons localised in the horizontal plane either around or in front of the user's head. By using head or hand gestures the user can select an auditory icon from the menu to trigger the corresponding event, for example, checking for traffic reports or weather. Billinghurst et al. [249] describe a wearable conferencing space using spatial audio to disambiguate between speakers. Kan et al. [178] present a laptop-based system using GPS to give spatial audio cues based on the actual location of the speakers relative to the listener.

More recently, Goose et al. [250] developed the Conferencing3 3D application that runs on PDAs. They combined VoIP software with spatial sound rendering and 3D graphics in a PDA client-server application. Spheres depicted on the PDA represent remote collaborators. The spatial audio is generated by a server PC, based on the position of the speakers relative to the user location in the conferencing space. The final audio stream is sent wirelessly to the PDA for playback. Deo et al. [251] showed how phone motion-tracking can be used to interact with mobile spatial audio content. Motion-tracking methods could be used to translate movement in the real world to orientation movements for navigating a virtual spatial audio space. Using phone and head tracking, Deo et al. conducted a user study evaluating these techniques in a spatial audio environment. They found that spatial audio modes using head and mobile phone tracking enabled better discrimination between speakers than fixed spatial and non-spatial audio modes. Spatialised audio with mobile phone orientation tracking provided the same level of speech intelligibility as head-tracking. Their study suggested that phone tracking is a viable option for orienting speakers in mobile virtual spatial audio environments.

The work summarised in this section is based around the smartphone form factor and its inertial tracking features. Given the imperative to develop non-visual interaction methods for navigation through complex interfaces, previous work has shown that gesture-tracking is an option worth exploring.

However, unlike most of the previous studies, the development of a generalisable and transferable set of gestures is desired. Existing concepts, such as the WIMP paradigm, are sought for inclusion in the gesture set but these have then been adapted to the requirements of a non-visual interface. As only an experimental verification can gain insight into the efficiency of the chosen approaches, in contrast to some of the related work mentioned above, a rigorous user evaluation was conducted.

7.2.3 Experimental Design

Given the model of a user surrounded by a sound space, the goal of this experiment is to determine which of two methods is best as measured by:

- time performance
- error rate
- task transitioning
- user preference

One of the novelties of the auditory display used is the translation into the auditory realm of the *foreground/background* metaphor that is much used in visual interfaces like Microsoft Windows or Apple OS. In these GUIs the user can minimize, maximize or tile windows depending on their focus of interest. Minimized windows may not deliver a constant stream of information, however they can be set to notify the user of a change of status, incoming messages, etc. Therefore, although users may have their main focus on a word processor they still have a sense of awareness of – in this case – their social network as represented by the messaging application. To support focusing of attention and monitoring of "background" events, an auditory interface was designed consisting of two concentric user-centric horizontal rings as illustrated in figure 7.5. By using the metaphor of *distance* to convey *importance*, the users were able to push items they are not currently focusing onto the outer ring, but this is achieved without depriving them of the option of monitoring and easily switching between different streams of information. It was of particular interest to observe participants using the phone itself as an input device to interact with sound sources in this 3D environment. More specifically, the research questions of interest were:

1. With no visual feedback, to what extent are users able to navigate and transition between several different audio streams of differing complexity?
2. In what ways do users utilise the perception of distance in the spatial sound environment to support task focus?
3. Which of the interaction methods did users prefer and why? How did these methods affect task performance with respect to time and degree of input required?

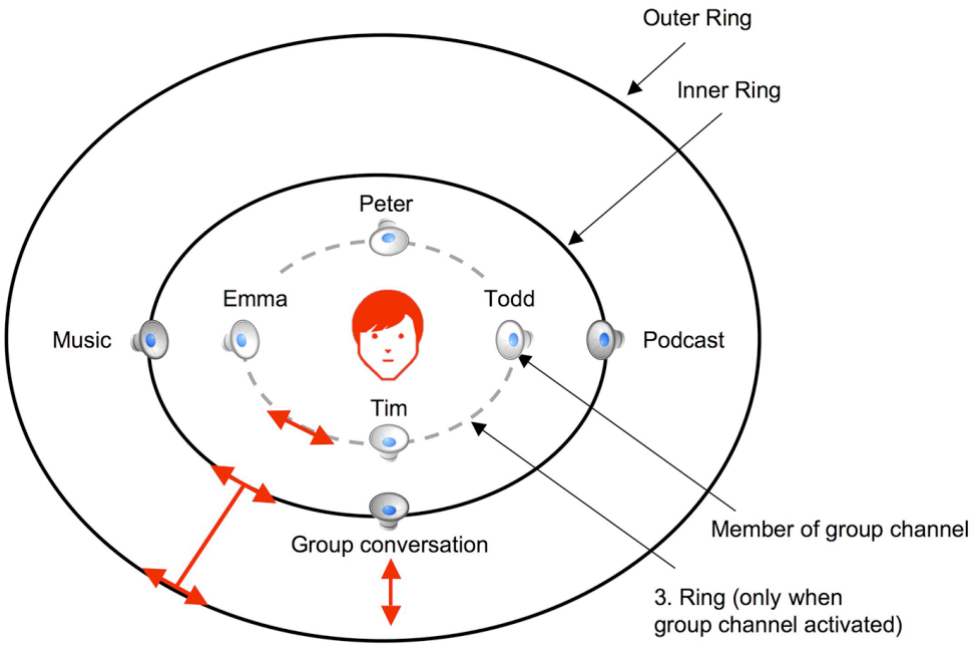


Figure 7.5: Layout of the soundspace, showing sound cues and interaction methods.

The conceptual model of a user surrounded by a sound space was used, and a prototype interface was developed to evaluate the experiment. The virtual spatial audio environment contained various types of audio items, including a simulated group conversation such as in a multiparty call.

The experiment used a standard computer, headphones, and a phone mock-up device (shown in figure 7.6). This enables enabled rapid application

development and testing on the PC. The phone device had an Intersense 3-DOF orientation sensor for motion direction tracking, in order to explore gesture-based interaction, and send tracking data via USB to the PC-based application.

Figure 7.6: The mobile phone mock-up device equipped with an Intersense Inertia Cube3 for motion tracking.

Interaction Metaphors: Push/Pull & Pan

The interaction metaphor for the experiment were push/pull and pan: Panning caused both rings (inner and outer) depicted in figure 7.5 to rotate clockwise or counter-clockwise. Items positioned on the rings rotated accordingly. A swooshing sound was played as feedback to a successful panning movement and the item in focus was announced. From the initial position rotating the rings counter-clockwise would announce "Music" to be in focus, then "Podcast", then "Group Conversation" then again "Music" and so forth. Furthermore, users could:

- Push the item in focus to the next farther ring (lower volume/farther distance)
- Pull the item in focus to the inner ring (louder volume/closer distance to user's head)

- Activate the item in focus by pulling beyond the inner ring

The sound source directly in front of the user – that is at 0 degrees azimuth, always had focus and was played at a slightly louder volume than the other items on the same ring. A sound source could only be moved out/in to a ring when it is had focus.

Figure 7.7: Panning gesture to rotate the sound scene and its items.

The two interaction methods compared in the study were: **Buttons**: the left and right button input was used to pan the rings counter-clockwise or clockwise; down to pull an item closer or to activate it, and up to push an item away or to deactivate it. **Gestures**: rotation of the phone left or right to pan the rings (shown in figure 7.7); vertical gesture upwards/towards the user for pulling the item in focus closer, and a vertical movement downwards/farther away from the user to push the item in focus away (shown in figure 7.8).

Audio Content

The audio content consisted of different types of audio to simulate both sporadic and continuous audio streams, group conversation, and system notifications.

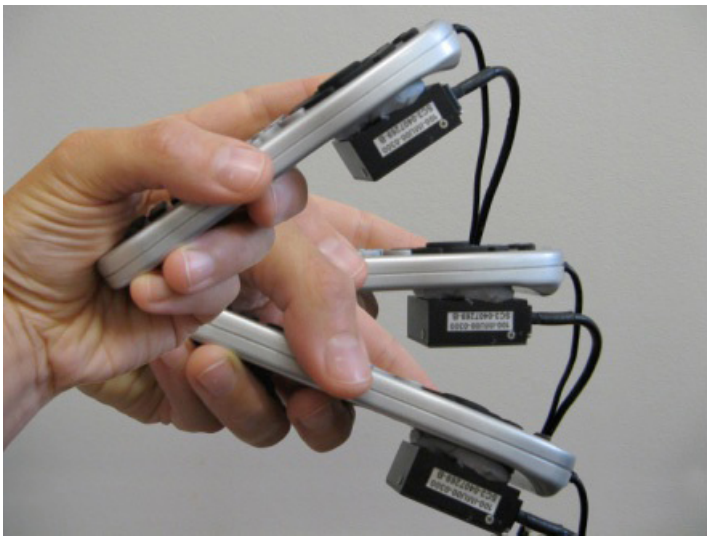


Figure 7.8: Pitch gesture to pull items closer or push them away.

The audio content used (as depicted in figure 7.5) was:

- Group channel (semi-continuous stream): four speakers having a pre-recorded conversation (sometimes overlapping in speech). Each speaker has a separate audio stream. The group-level (combined) audio was initially placed directly in front of the user, on the inner ring.
- Music (continuous stream): to the speaker's right at 90 degrees.
- Podcast (continuous stream): individual audio stream representing speech-based audio content of interest to the user. Initially placed to the user's left at 90 degrees.
- Notification beeps (sporadic audio items): representing generic events, such as incoming phone call, or calendar event. These occurred at random intervals in the final task only (see subsection 7.2.3 for a description of the tasks.). They were played in stereo and were not spatialised.

Transitioning mechanism

A secondary goal of this study was to gauge the efficacy of the transitioning mechanism between individual speakers and a group of speakers. The system was designed to allow listeners to identify, isolate, and directly communicate

with an individual from a group. One individual speaker audio channel playing a Podcast, and one four-speaker "synchronized" group-level channel, in which a pre-recorded group conversation was taking place, were included. The transitioning mechanism worked as follows: If the group-level channel was located on the inner ring as illustrated in figure 7.5, the pitch gesture, i.e. pulling the group channel closer, created a temporary ring composed of the group's individual speakers, spatially separated around the user. While the group channel was activated, the Podcast and music were muted and the user could hear only the group conversation. Activating the Podcast or the music by pulling the item from the inner ring towards the user muted all other sounds. While activated the sound source had no spatial effect but was played in stereo. All sounds could be deactivated by pushing them away (onto the inner ring), which restored the previous layout.

The group members were, by default, played at the same volume level. In order to isolate a particular member, the user had to rotate the ring until that speaker was at the 0 degrees azimuth position. In order to "whisper" (that is, communicate on a dedicated channel) to that person, the user had to pull them closer. While whispering to a person all other group members were muted.

Experimental Apparatus

The hardware set-up consisted of a Mavael Keiboard equipped with an Intersense Inertia Cube3 (IC3) for 6-DOF gesture tracking (see figures 7.6 and 7.8). To guarantee an authentic distance effect, each individual sound stream was pre-recorded very close to the speaker and at the same time from a distance of about three meters. For positioning the virtual sound sources were processed with the application of HRTFs from the fmod sound library[7]. A pre-study showed that the distance effect in the recorded files was not perceived to be very distinct. Instead, paying particular attention to preservation of intelligibility, attenuation and reverberation were post-processed to achieve a clear, recognizable distance effect.

[7] `http://www.fmod.org`

Figure 7.9: Experimental setup.

All four speakers of the group conversation were recorded separately. These four tracks were initially played from the same position. If the group conversation was activated these four tracks were spatially fanned out as can be seen in figure 7.5. Special care was put into the recording procedure to ensure authentic reproduction of each speaker's individual characteristics as Yankelovich et al. [110] have shown that audio quality has a strong influence on the effort required to understand the meaning of sentences and on the perceived sense of presence in a teleconferencing environment. During the testing users wore Sony MDR-V700 adjustable headphones.

Experimental Procedure

The study design was a within-subjects experiment where all participants solved tasks using both interfaces, namely input via buttons on the keypad and using motion tracking of the phone for gesture input.

The dependent variables were:

- task completion time
- number of interactions
- number of missed notifications (only for task 4)

Additionally, the results of an extensive post-study questionnaire on user interaction satisfaction, perception of sounds and sound localization were evaluated. 34 people participated in the study. The mean age was 33.9 years spanning from 16 to 57 years, and approximately half of the participants were below thirty years of age. 47 percent of the participants were female, and 53 percent were male. The majority regarded English as their first language (70.6 percent). The majority of the participants (73.5 percent) reported using a computer more than 30 hours per week. 32.4 percent of the participants regarded themselves as quite experienced mobile phone users, with another 23.5 percent regarding themselves as expert. Just over half of the participants felt themselves to be novices with auditory interfaces (58.8 percent). Two of the participants reported to have minor hearing difficulties but were not excluded from the study because of the negligibility of their impairments.

Participants were presented with each of the four tasks in the same order, but with an alternating order of the interaction methods. Before starting the tasks participants were asked to familiarize themselves with the technology until they felt to have a good understanding of the interface and the interaction methods. Figure 7.10 depicts the interaction device and the participant information sheet available to users during the study.

Tasks

Each subject was presented with the following four tasks, in the given order:

T1: Please move the music to the position 0 degrees azimuth (right in front of you) and push it to the outer ring.

T2: You want to monitor all sounds but the group channel is distracting you. What do you do?

T3: You would like to concentrate on the group conversation. What do you do?

T4: Please identify the member of the group channel who is still working. Once you've done so please open a whisper/private channel to this member by pulling him or her closer. While you are doing so, please hit SPACEBAR whenever you hear a notification.

T4 was designed to be the longest and most complex task. The number of notifications given, and the number of these responded to by pressing the spacebar was also logged. After data was collected for each condition subjects filled out a subjective survey giving responses on a number of questions such as how easily they were able to navigate between audio streams, how intuitive each interface was, which interaction method they preferred, as well as whether they would consider using the application for group-based communication in daily life.

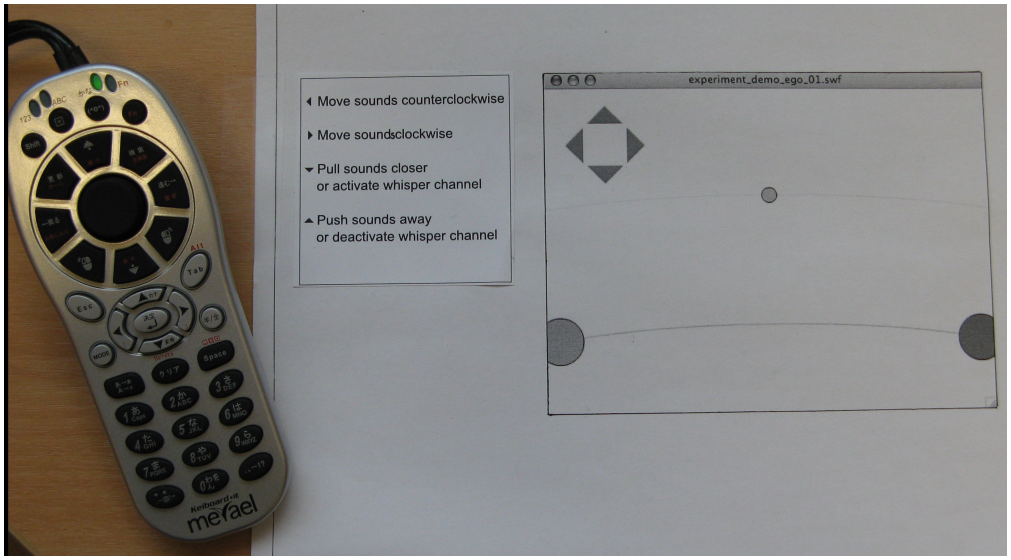


Figure 7.10: Interaction device and information sheet available to participants during the warm-up and test phases.

7.2.4 Results

The following section summarizes the analysis of the gathered data. The performance of both interaction methods is evaluated in terms of the time and effort it took to complete tasks, and, in the case of task 4, the number of notifications missed. As the data was normally distributed and scale-levelled, unless otherwise stated, a *paired t-test* with a fixed confidence level (p-value = .05) was used to analyse the data. Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the data set. Furthermore, the results of the post-study questionnaire are reported.

Performance with Buttons vs Gestures

Any of T1, T2, and T3 could be solved with only two interactions while T4 required at least three interactions. An interaction is either a left/right/up/down press of a button on the keypad or a lateral/pitch movement of the phone in the case of gestures. Figure 7.11 shows the mean number of interactions per task and figure 7.12 depicts mean task completion times.

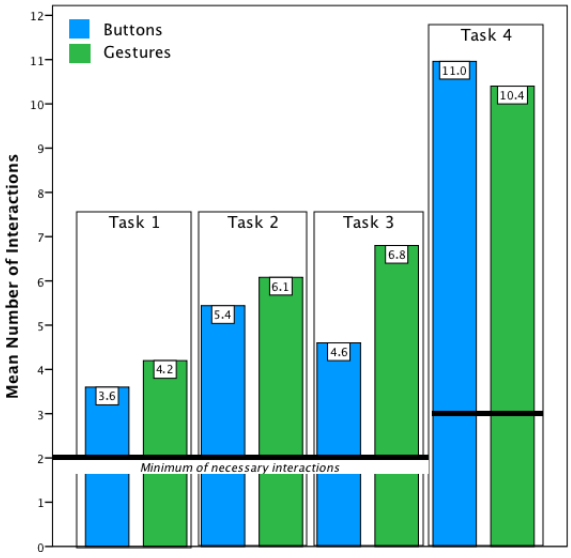


Figure 7.11: Mean number of interactions in both conditions and over all tasks. The black line marks the number of minimal required interactions.

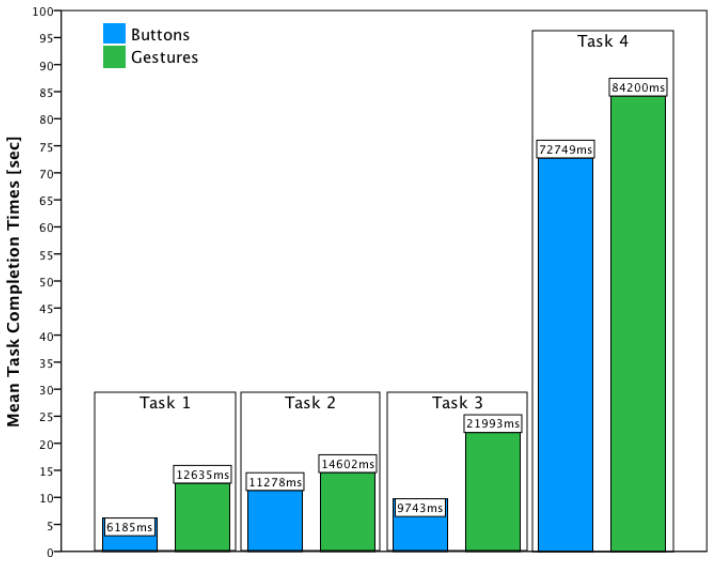


Figure 7.12: Mean task completion times for both conditions and over all tasks.

Table 7.2 gives an overview of the performance, including mean task completion times and interactions of participants' when using buttons. Table 7.3 lists the corresponding overview for participants' using gestures. Table 7.4 presents the strength of the correlation (using Pearson Correlation coefficients) between interactions made and time for task completion, across both conditions and all four tasks. There were – as expected – moderate to strong correlations found between the number of interactions made and task completion time across the conditions.

For T1 no significant difference ($t(30) = 0.16$, $p = .88$) was found in the number of interactions for buttons ($M = 4.2$, $SD = 3.5$) and gestures ($M = 4$, $SD = 2.5$). This is also the case for task completion times ($t(30) = -1.89$, $p = .07$).

Task	**Task Completion Time [msec]**		**Interactions**	
	N	M, SD	N	M, SD
1	32	M = 6911 SD = 8248	32	M = 4.1 SD = 3.4
2	32	M = 20820 SD = 41660	31	M = 6.8 SD = 7.0
3	32	M = 12065 SD = 14098	30	M = 5.2 SD = 3.5
4	30	M = 74748 SD = 29527	32	M = 10.6 SD = 6.9

Table 7.2: For condition **Buttons**: Number of valid N, task completion times, and mean number of interactions per task.

Task	**Task Completion Time [msec]**		**Interactions**	
	N	M, SD	N	M, SD
1	33	M = 13256 SD = 12763	32	M = 4.0 SD = 2.5
2	32	M = 16481 SD = 19265	31	M = 6.3 SD = 5.1
3	33	M = 22680 SD = 27161	31	M = 7.0 SD = 4.4
4	32	M = 81950 SD = 35532	31	M = 10.1 SD = 6.1

Table 7.3: For condition **Gestures**: Number of valid N, task completion times, and mean number of interactions per task.

Task	**Buttons** Correlations between interactions and task completion time	**Gestures** Correlations between interactions and task completion time
1	$r = .83$, $p < .01$	$r = .68$, $p < .01$
2	$r = .97$, $p < .01$	$r = .79$, $p < .01$
3	$r = .76$, $p < .01$	$r = .68$, $p < .01$
4	$r = .46$, $p < .05$	$r = .57$, $p < .01$

Table 7.4: Correlations between the number of interactions and task completion time for both interaction techniques per task.

For T2 all participants except for two chose between two different approaches (A1 or A2). To recap, T2 involved monitoring all sounds whilst finding that the group channel was distracting, and therefore finding a way to deal with this issue. The approaches in response to this task were:

A1: Pull music back onto inner ring, push group conversation onto outer ring, chosen by 17 participants when using buttons and 20 when using gestures.

A2: Leave music on outer ring, push group conversation onto outer ring, chosen by 13 participants when using buttons and 11 when using gestures.

An independent-samples t-test showed no significant difference regarding the task completion time ($t(28) = 1.12$, $p = .27$) and the mean number of interactions made ($t(28) = 1.27$, $p = .27$) across the approaches when using buttons. Similarly, no significant differences were found when using gestures regarding task completion times ($t(28) = .95$, $p = .46$) and interactions ($t(28) = .89$, $p = .69$).

T3 involved concentrating on the group channel. The approaches in response to this task were:

A1 : Pushing music and Podcast onto the outer ring, bringing group conversation onto the inner ring (this allowed monitoring of the two other sound sources while focusing on the group conversation), chosen by 20 participants when using buttons and 21 when using gestures.

A2 : Pulling group conversation to the inner ring, then activating group conversation (which muted all other sound sources), chosen by 13 participants when using buttons and 11 when using gestures.

For T3, no significant difference in the task completion time ($t(30) = -0.78$, $p = .44$) were found using either approach. However, in comparing the number of interactions made on average between the approaches for this task, those who used approach A1 made more interactions when using buttons ($t(28) = -2.33$, $p = .03$). The same is the case for gestures: No significant differences in task completion times ($t(30) = -1.96$, $p = .06$) were found but there were differences in the number of interactions made (($t(29) = -2.4$, $p = .03$). On this basis, approach 1 appears to have necessitated more by way of input from the users without affecting their task completion times significantly.

For T4, no significant differences between the techniques, either in terms of task completion times ($t(28) = -0.69$, $p = .5$) or the number of interactions ($t(30) = .53$, $p = .06$), were found.

Sequence Effects

When subjects commenced the study using the gestures interaction method, they rated using buttons ($M = 4.5$, $SD = .6$) significantly more highly ($t(32) = 2.55$, $p < .05$) for straightforwardness of use than participants who started by using buttons ($M = 3.9$, $SD = .8$).

The sequence in which tasks were presented also had an impact on completion time and the number of interactions made. Participants were faster and used fewer interactions in both T2 and T4 when their first condition was gestures. For T2, using buttons was significantly faster ($M = 7.4$, $SD = 5.6$, $t(30) = -2.04$, $p = .05$) and significantly fewer interactions were made ($M = 4.5$, $SD = 2$, $t(29) = -2.06$, $p < .05$) when participants started with gestures than participants starting with buttons (task completion time: $M = 36$, $SD = 57.8$; interactions: $M = 9.5$, $SD = 9.8$).

For T4, the same learning effect could be observed. Again, for the condition "buttons" participants were significantly faster ($M = 63.6$, $SD = 12.8$, $t(28) = -2.81$, $p < .01$) and used significantly fewer interactions ($M = 8.3$,

SD = 5.1, t(30) = -2.21, $p < .05$) when they started with gestures. Starting with buttons had a negative effect on the mean scores for the "button" condition as participants were slower (M = 91.4, SD = 39.3) and made more interactions (M = 13.4, SD = 7.9).

Participants performing in the sequence of first using gestures and then using buttons to complete the tasks showed a strong learning effect for solving tasks by pressing buttons. This could be due to knowledge and experience gained while using gestures. The same effect could not be observed for using buttons first and then gestures: it may be the case that the observed difficulties some participants had with the gesture recognition forced them to concentrate more on the interaction technique itself which may have cancelled out the learning effect.

Missed Notifications

For T4, participants were asked to perform a secondary task of listening for system notification beeps, conceptually representing the arrival of emails or incoming calls, whilst conducting the primary task of identifying a particular group member and whispering to them. On average, users missed .6 notifications, or approximately fifteen percent of the notifications that each participant heard. No significant difference (t(30) = -.23, p = .82) could be found for missed notifications between the two interaction conditions.

7.2.5 User Satisfaction

Tables 7.5 and 7.6 present a summary of participant responses[8] to the post-study questionnaire. In the questionnaire the interaction technique were compared on several levels including navigability between sounds sources, straightforwardness of use (thus aiding ease of learning), and the overall user preference. Also, the overall satisfaction with the interface design was evaluated.

[8] On a Likert Scale of 1 equalling the negative maximum and 5 equalling the positive maximum rating.

Questions included:

Are you satisfied with the accuracy of the system?

Not at all < 1 - 2 - 3 - 4 - 5 > very much so

Is the system easy to use?

Not at all < 1 - 2 - 3 - 4 - 5 > very much so

Using the application is/feels:

difficult < 1 - 2 - 3 - 4 - 5 > easy

frustrating < 1 - 2 - 3 - 4 - 5 > satisfying

dull < 1 - 2 - 3 - 4 - 5 > fun

Interface Aspect	**Buttons, % pos. response**	**Gestures, % pos. response**
Learning to navigate the system is easy.	73.5	65.2
All given tasks can be performed straightforward.	85.3	52.9
My location within the system at any given time is apparent.	67.7	64.7
I am satisfied with the accuracy of the interaction technique.	100	61.8
I liked using buttons better than gestures.	61.7	
I liked using gestures better than buttons.	32.7	

Table 7.5: Participants' ratings of interaction methods.

Ratings of 1 or 2 were grouped as negative responses, 3 was regarded as undecided, and 4 and 5 as positive responses. The results for these preference ratings can be summarized as follows:

In learning to navigate the interface, buttons ($M = 4.3$, $SD = 1$) were deemed to be significantly easier ($t(33) = 3.42$, $p < .01$) than using phone gestures ($M = 3.5$, $SD = 1.2$).

On straightforwardness of use, buttons were rated to be significantly more straightforward ($t(33) = 3.32$, $p < .01$). The average rating for buttons was $M = 4.2$ ($SD = .8$), and for gestures $M = 3.6$ ($SD = 1.1$). In terms of the accuracy of the interaction technique, buttons were deemed significantly more accurate ($t(33) = 4.51$, $p < .01$), with average scores of $M = 4.6$ ($SD = .5$) for buttons and $M = 3.5$ ($SD = 1.4$) for gestures.[9] Overall, buttons ($M = 3.8$, $SD = 1.3$) were preferred over gestures ($M = 2.7$, $SD = 1.3$) as interaction method for this application ($t(33) = 2.71$, $p = .01$).

There was no significant difference found in sense of location perceived within the system, as afforded by either technique. Table 7.6 presents the percentage of positive and negative reactions by participants to various aspects of the system as a whole.

Item	**% Pos. Resp.**	**% Undecided**	**% Neg. Resp.**
Auditory Display Design			
System Accuracy	79.2	-	20.6
Auditory Nature	67.7	20.6	11.8
Audio Layout	91.1	5.9	2.9
Easiness of Usage	76.4	17.6	5.9
System Efficiency	72.7	18.2	9.1
User Interaction Satisfaction			
Terrible – Wonderful	64.7	32.4	2.9
Difficult – Easy	70.6	5.9	23.5
Frustrating – Satisfying	61.7	32.4	5.9
Dull – Easy	91.2	5.9	2.9

[9] However, when transferring these results to other gestural or tactile interaction techniques, it should be borne in mind that the implementation was prototypical and not a fully developed off-the-shelf solution. Therefore, when considering these results the prototypical implementation of the gestural interaction device ought to be considered as a confounding factor, which may have had an impact on the ratings.

Slow – Fast	52.9	38.2	8.8
Boring – Stimulating	82.4	14.7	2.9
Impersonal – Personal	70.6	20.6	8.8
Passive – Active	82.3	8.8	8.8
Focus & Distraction			
Support for group awareness	58.8	26.5	14.7
Aids concentration on other tasks	29.4	26.5	44.1
Causes distraction from other tasks	55.8	20.6	23.5
Aids monitoring other tasks	67.7	20.6	32.4
Aids connectedness to social network	70.6	17.6	11.8
Sound Quality & Spatiality			
Sound Identification	70.6	20.6	8.8
Overall Quality of Sound	91.2	8.8	0
Sound Position Identification	64.9	20.6	13.5
Helpfulness of Spatial Sound	85.3	11.8	2.9
Distance Effect of Sound	67.7 (good)	20.6	11.7 (poor)

Table 7.6: Participants' interaction satisfaction responses.

In addition, in response to questions about whether they would use this type of application for group awareness activities in daily life, 8.8 percent said never, 50 percent said occasionally, 32.4 responded often, and 8.8 percent responded always.

7.2.6 Gender Effect and Lab Affiliation

The results for the overall rating of the interface were strongly influenced by either gender or affiliation with the laboratory. The source of the influence cannot be derived with absolute certainly as most male participants (all except for three) worked at the laboratory and most female (all except for four) did not. The variables representing gender and affiliation with the laboratory show a significant correlation ($N = 34$, $r = .589$, $p < .01$).

The results show no significant difference between genders/affiliation with the laboratory measured on task completion time, number of interactions made, and missed notations. But for task T3, (using gestures) a significant relationship χ^2 (2, $N = 32$) $= 9.219$, $p < .01$) between men/members and approach 2 was found. That is, men/members ($N = 15$) expanded the group call and therewith muted all other sources more often than women/non-members ($N = 6$). Nine women/non-members chose approach A1, which means they brought the group conversation to the inner ring and moved all other sound sources to the outer ring, in comparison, only two men/members did the same. No significant effect could be found for T3 when using buttons.

Table 7.7 summarizes results from an *independent-samples t-test* on general satisfaction with the interface.[10] It shows that ratings from women/ participants not from the laboratory were significantly higher for some items on the questionnaire. Observations during the study would rather support the interpretation that the actual influential factor is affiliation with the HIT Laboratory NZ. Participants who were not from the lab were more excited about the experiment itself and about using the interface. As they were not as experienced in dealing with new technologies and multimodal prototypes as participants from the laboratory, the uniqueness of the whole experience might have influenced the ratings. On the other hand, research suggests that women have better hearing at frequencies above 2000 Hz (frequency range of speech is approx. between 150-5000 Hz) [252, 253].

[10] Results are responses on a 5-point Likert Scale with 1 representing the left word of the pairing (e.g. frustrating) and 5 the right word of the pairing (e.g. satisfying).

	p	**t**	**Mean** Women/Non-Members	**SD**	**Mean** Men/ Members	**SD**
Accuracy of System	<.01	(32), 3.37	4.7	0.6	3.9	0.8
Easiness of Usage	.01	(32), 2.76	4.5	0.7	3.7	0.9
Efficiency of System	<.01	(31), 3.09	4.6	0.6	3.7	1.1
Respond in Real Time	<.01	(31), 3	4.6	0.6	4	0.6
Frustrating – Satisfying	<.01	(32), 3.04	4.2	0.9	3.4	0.6
Dull – Fun	<.01	(32), 3.09	4.7	0.5	4	0.8
Boring – Stimulating	<.05	(32), 2.45	4.4	0.5	3.8	0.9
Insensitive – Sensitive	<.05	(32), 2.59	4.3	0.8	3.4	0.8
Cold – Warm	<.01	(32), 2.84	4.1	0.7	3.4	1.1
Passive – Active	<.01	(32), 2.62	4.7	0.5	3.9	1.1
Aids monitoring tasks	<.05	(32), 1.77	4.1	1	3.6	0.9
Sound Position Id.	<.05	(32), 2.1	3.7	1.1	3	1.3

Table 7.7: Significant differences in the post-study questionnaire data between women and men or participants who were recruited from outside the laboratory (non-members) or among the affiliates of the laboratory (members).

Also, women are more likely to engage in the elaborative processing of the meaning of verbal (or verbally encoded) information [254]. These factors suggest that it may have been easier or more enjoyable for women to operate the system. On the basis of the results discussed above, the need for further research of these variables is clearly apparent.

7.2.7 Discussion

It appeared that the gestural interaction technique initially produced more errors and confusion. The novelty of this interaction technique is reflected in the results where users who commenced the experiment using gestures rated using buttons as more straightforward than users who started with buttons. Buttons also fared better in participants' estimation of the usefulness of the system for social networking. For T2 and T4, task completion times and number of interactions made were also affected by this sequence effect with

buttons taking less time to master and fewer interactions when preceded by gestures in the sequence of task testing. It can, therefore, most probably be concluded that it took participants longer to form an adequate mental model if they began the study using gestures. Some participants commented that they would have preferred to have more feedback for the pitch interactions, corresponding to the level provided with the lateral interactions (for which a "swooshing" sound was heard), and that it consequently took them more effort to develop their conceptual model of the system.

When participants started with buttons no effect of sequence was observed. This may be due to the more rapid formation of a correct mental model without the confusion deriving from the unfamiliarity of gesturing with the phone as an interaction technique. Gaining a good understanding of the interface while using buttons may have compensated for irritations produced when using gestures to interact with the system. However, it was observed that once participants understood how to operate the interface they were soon became much faster and more precise (that is, on the basis of making fewer unnecessary interactions). Thus it appears that there was a distinct learning effect when buttons were used first.

On the basis of the secondary listening task in T4, whereby participants monitored system notifications and responded to them, the interaction method used was not found to affect participants' response rates. Thus, it could be argued that using gestures did not produce a higher workload than using buttons.

7.2.8 Conclusion

The analysis of this study into the usability of eyes-free interaction techniques, spatial sound and the metaphor of *distance*, provided interesting insights with regards to the initial research questions: **RQ 2**: What are viable non-visual multimodal interaction techniques? **RQ 2.1**: What are the advantages and disadvantages of different tactile interaction techniques? and **RQ 5**: How can the focus of attention be supported?

These findings are summarized below.

1. Interaction Techniques

- While most people preferred using directional keys on the device for this application, approximately one third of participants preferred using gestures. Users' comments suggest that they liked the playfulness and intuitiveness of the gesture-based interaction method. These findings suggest a positive pathway for the acquisition of more familiarity with the gestural interaction technique.
- No visual feedback was given in this study. Participants successfully used auditory feedback as well as the their own kinesthetic awareness through gesture movement, to aid in imparting a sense of the virtual three-dimensional sound environment.
- User satisfaction ratings with the overall system were very high, particularly with respect to the elements of fun, stimulation, and active participation. Users stated that they had no problems identifying single sound items. Item selection and manipulation could be easily accomplished.

2. Focus of Attention

- By enabling participants to pursue their own approach to T2 and T3, it was discovered that some participants embraced the *distance* metaphor as a means to focus their attention on one sound source. An optimistic appraisal of distance as a means to support multitasking and the focus of attention in three-dimensional audio interfaces might be well justified.
- Approaches chosen for T2 and T3 indicate that applying a distance effect is a viable option to support background awareness and focus direction.

3. Mental model, awareness of system state, awareness of location within the system

- There were negligible rates of failure to complete the tasks. Once participants felt acquainted with the application they had a correct understanding and a good sense of their location within the system. Objects could be correctly identified and successfully manipulated.

- Almost all participants had a good sense of the spatiality of the environment independent of the interaction techniques. Enhancements to the spatial effect, like for example adding reverberation effects, and additions to the proximity feedback of the sound sources are likely to further improve the system.

Overall, the results of this study were positive for the continuation of further exploration of multimodal interaction for purely auditory interfaces as a complement to or a substitute for visual interfaces. Responses were strong regarding the applicability to group communication and multitasking. It seems fair to assume that the proposed interface would also be applicable for other systems which support multitasking and require the focus of attention, such as mobile phones, music players, or digital assistants and also assistive technology for the visually impaired.

The results also point towards an optimistic estimation for utilising spatial sound to promote a feeling of greater connectedness with social networks. Currently, most social network technology is heavily based on visual cues. Offering a constant but only sporadically utilised audio connection, as simulated in the experiment, may be a viable and less disruptive alternative for supporting group communication, awareness, and a feeling of social presence in both real and virtual social networks, especially in mobile scenarios. These seem to be very fruitful areas for future research.

7.3 Empirical Study 2: The Impact of Two Eyes-free Interfaces On a Demanding Primary Task

The first two sections of this chapter focused on addressing how a user can intuitively interact with an eyes-free interface and how this interface could be designed to support multitasking and the focus of attention. Based on the lessons learned in the previous studies, the study described in this section explores the impact on primary task performance, similar to Task 4 in section 7.2.3 but in a much more realistic "mobile" scenario. For this purpose, two auditory interfaces were compared to a visual head-down display for navigating a mobile phone menu while driving. Unlike in the previous study 7.2, this study had a much more challenging primary task, i.e. participants were asked to steer a vehicle while solving secondary tasks on the simulated mobile phone interface. Using the phone itself as the input device made sense in the two prior experiments, but encouraging drivers to take their hands off the steering wheel is not advisable due to the haptic distraction it would cause. Therefore, a customized input device was built for this study. It consisted of two mouse buttons, and a scrolling wheel attached to a steering wheel (see figure 7.15). As the auditory interface built for the first empirical study proved to be a user friendly, efficient, and conceptually flexible approach, the auditory interface designed for this study inherits many key features of the original design; namely the circular layout, selection and navigation methods.

Before presenting the interfaces and experimental results in more detail, the following section briefly summarizes the significance of attention in the context of driving a vehicle. For an overview of related work on auditory interfaces and especially those using a circular arrangement of sound sources, the reader is referred to section 2.6.4 in chapter 2. In section 7.3.2 the design rationale is reflected upon and a detailed description of the experimental design is given (7.3.3), followed by sections on the user study (7.3.4) and its results (7.3.5). This section concludes with a discussion (7.3.6) and a recap of the study, its findings, and some design recommendations in the conclusions section (7.3.7).

7.3.1 Related Work on Attention & Distraction

One of the central issues of interacting with interfaces other than operational controls of the vehicle is that, according to the Multiple Resource Theory of Attention (see section 2.2 for an elaboration), attention is diverted away from the primary task, which can have a critical impact on driver performance. Such a diversion, e.g. when a driver talks on a mobile phone or interacts with a navigation system, is more formally defined by the AAA Foundation for Traffic Safety [255] as:

> "[...] a driver is delayed in the recognition of information needed to safely accomplish the driving task because some event, activity, object, or person within or outside the vehicle compelled or tended to induce the driver's shifting attention away from the driving task."

Distraction, which is distinguished from inattentiveness by the presence of a triggering event, is not only caused by physical stimuli through the sensual apparatus, but also by cognitive sources, such as thought or emotional arousal [256, 257]. In multitasking situations it leads to a reduced amount of attention on either task, the initial or primary and the new or secondary task [59].

Distraction from the primary task, i.e. driving the car, can reduce driver safety by degrading the vehicle control, such as speed maintenance and lane keeping, and also degrading object or event detection [258]. Apart from visual (*eyes-off-the-road*), auditory, and cognitive distractions (*mind-off-the-road*), mechanical causes can also distract as drivers who are reaching for objects inside the vehicle or are otherwise shifting out of their normal sitting position can have a degraded ability to execute manoeuvres [258, 56]. Therefore, handling a mobile phone while driving a vehicle or being otherwise on the move differs fundamentally from using it in the office or at home.

In this study, the problem of interfering sensory resources has been investigated by comparing two auditory to one visual interface for navigating the menu of a mobile phone. To reduce the amount of mechanical distraction, the physical interaction device was attached to the steering wheel, so the driver's hands could remain on the steering wheel. However, Llaneras [259]

points out that although visual and mechanical distraction can be partially reduced in this way, cognitive distraction is not eliminated.

7.3.2 Design Rationale

The design rationale of this study was to investigate a) how distracting an eyes-free interface similar to that described and evaluated in 7.2 is in the context of a realistic primary task and b) how the eyes-free alternative performs in comparison to a standard visual interface. To simulate a case of realistic usage, the experiment was run in a driving simulator. Participants performed tasks of differing complexity while they drove the simulated vehicle. The driving simulator, which is depicted in figure 7.16 consisted of a large projection screen, steering wheel, accelerator, brake and mobile communication device which could be controlled with a custom-made interaction device attached to the steering wheel (see figure 7.15). In addition, an external keyboard was attached next to the steering wheel, which could be used for entering letters or text if necessary. A more detailed description of the experimental apparatus is given in section 7.3.4.

Figure 7.13: The visual interaction based on a small screen and phone-like keyboard. The items in the visual menu were displayed in large white fonts and the selected item was highlighted with a green bar.

7.3.3 Experimental Design

Three different user interfaces, one visual and two auditory interfaces, were compared in the study. The same menu structure was used with all three interfaces. The items and the levels of the menu were based on a Nokia 60-series mobile phone menu but were reduced to a set of items most likely to be accessed in a mobile situation. There were up to six items on each level with the top level containing the following items:

- Messaging
- Contacts
- Gallery
- Media
- Profiles
- Tools

Condition 1: Visual Interface (V)

Figure 7.13 depicts the visual interface used for the experiment. The screen was positioned at about 40 degrees to the lower left of the dashboard where it could be easily seen while driving. The menu items were displayed in large white fonts and the selected item was highlighted with a green bar. By pressing the left mouse button attached to the steering wheel, the user could descend through the menu's hierarchy. By clicking the right mouse button they could ascend. In the tasks where a text message had to be entered, the small phone-like keyboard was used for entering individual letters.

Condition 2: Auditory Interface with multiple sounds playing (AM)

In the second condition (AM) all menu items and commands were presented to the driver via the loudspeakers installed in the simulator. Figure 7.17 shows a layout of the loudspeaker setup. As speech has proven to be very effective in auditory interfaces [138, 139], the same menu items presented on the screen in the visual interface were read by a native English speaker, pre-recorded for the experiment, and used in both auditory conditions. Initially, each of the six menu items was placed on a virtual circle and assigned to

one of the virtual six sound sources. By using the scrolling wheel attached to the steering wheel, the circle – and with it the sound sources – could be rotated around the user's head. In condition AM all menu items of the current level of the menu were played simultaneously. Participants could select an item by rotating the circle until the desired item was placed directly in front of the user at 0 degrees azimuth (see figure 7.17). To emphasize the "selected" object, items positioned on the front loudspeaker were increased in volume. By presenting all items of one menu level simultaneously, users are given an overview of all items available. However, this may be at the expense of reduced intelligibility due to masking effects (see section 2.1.4 for an introduction to auditory masking.). To aid orientation within the menu structure, a gentle background melody was assigned to each individual branch of the menu. The melody started as soon as the user left the main menu and entered one of the sub-menus. The central pitch of the melody was changed according to the current depth of the user with in the sub-menu structure. Each time the user moved to a lower level of the menu, the pitch was lowered and vice versa.

Condition 3: Auditory Interface with a single sound playing (AS)

In the third condition (AS), the positions of the menu items – the sound sources – were the same as in the AM condition, but with the difference that only the front sound source was played. Conceptually, this interface behaves like an invisible/inaudible ring that is rotated around the user. A slot which allows the item on that position to be seen/head occurs in front of the user only.

In both auditory conditions (AM and AS), textual input was also realized with an acoustic interface. Two major letter groups (vowels and consonants) were represented on one level of the menu together with *Space*, *Erase letter* and some other commands (see figure 7.14). On the next level, the consonants were further divided into six smaller groups of letters, such as {b, c, d} and {f, g, h}. Below that level each single letter was represented and could be selected. After each selection of an individual letter the user was automatically moved back to the first level of text input menu. For example,

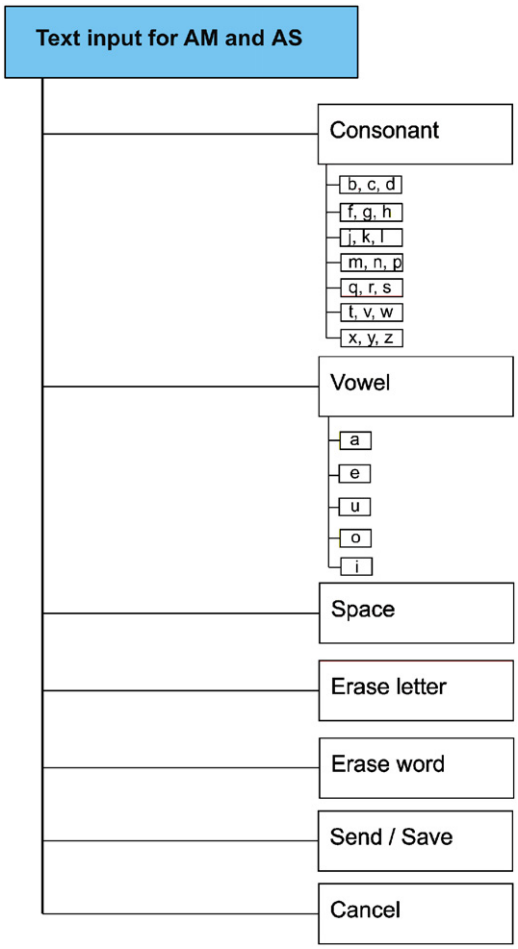


Figure 7.14: A diagram of the auditory menu used to compose text messages in the AM and AS conditions.

to compose a text message “HI”, the user would first need to select the group “Consonant”, then the group “f, g, h” and then the letter “h”.

After this selection, the user would be automatically moved back to initial position of the text input and would again need to select between "Consonant", "Vowel", etc. This time the user would select the group "Vowel" and then further the letter "i". The input of the message "Hi" would thus be completed (see Fig. 7.14).

Interaction Technique

In all three cases the interaction in the experiment is conducted through a custom-made interaction device consisting of a small scrolling wheel and two buttons (left and right) attached to the steering wheel (see figure 7.15). In the first condition (V), the scrolling wheel was used to move the selection bar up and down in the menu and the two buttons were used to confirm or cancel the selection. In the second (AM) and third (AS) condition, the scrolling wheel turned the virtual circle with the sound sources, thus changing the position of the items in the menu. The item in front of the user was always automatically selected. As in the visual condition, the two buttons again enabled the user to enter the sub-menu behind the selected item (left button) or ascend to the next higher level (right button).

Tasks

It was of interest to observe the users operating the car (primary task) and at the same time performing different (secondary) tasks with the in-built mobile device. Participants were instructed to put more emphasis on safety than on speed in the driving task. Driver performance was measured according to the number of errors made, such as lateral deviations, reducing speed significantly when it was not required, driving off the side of the road or even crashing the car. Participants were asked to perform five different secondary tasks while driving. These were:

1. Writing a text message to a specific person (MSG)
2. Changing the active profile of the device (PRF)
3. Making a call to a specific person (CAL)
4. Deleting a specific image from the device (IMG)
5. Playing a specific song (SNG)

The main research questions were:

1. Which interface will distract the user less from the primary task?
2. Which interface will cause the user to make more errors?
3. Which interface will have the shortest task completion times?
4. Will the audio interface with multiple simultaneous sounds (AM) be more distracting than the audio interface with just one sound (AS)?

Hypotheses

Due to the Multiple Resource Theory of Attention it was expected that the use of the auditory interfaces (AM and AS) would distract the users less from the primary, mostly visual, driving task than the visual interface (V). Consequently, the driving performance should therefore be significantly better in conditions AM and AS. For the same reason shorter task completion times for AM and AS were expected, especially with simple tasks, such as changing the user profile or calling someone.

Figure 7.15: The interaction device consisting of a scrolling wheel and two mouse buttons.

Comparing AM and AS, the AM condition was expected to be more efficient due to a larger information flow as many sounds were played simultaneously. Users should therefore have a better awareness of their current position in the menu and the positions of individual items in the levels of the menu. However, the simultaneous playback of many items may be perceived to be noisy and confusing and therefore participants may prefer the AS condition where only one sound is played at a time.

7.3.4 User Study

A total of 18 test subjects (8 female and 10 male) participated in the experiment. The average age of the test subjects was 27.7 years with an average of 8.7 years of driving experiences. Half of the test subjects were more experienced with driving on the left-hand side of the road and half of them on the right-hand side. They all reported normal sight and hearing.

In order to eliminate learning effects between the different interfaces, a within-groups design was chosen for the study. Three groups of six participants were formed. Each group performed the tasks in a counterbalanced sequential order:

- Group 1: V, AM, AS
- Group 2: AS, AM, V
- Group 3: AM, V, AS

An additional group of five test subjects (1 female and 4 male) participated as a control group without performing any secondary tasks. In all conditions the test subjects were asked to drive the car safely and perform the tasks as quickly as possible. Each task was read to the test subjects loudly and clearly. The tasks were ordered randomly for each interface. The successful completion of the individual tasks was signalled with the message "Task completed" either on the screen or played through the loudspeakers. The duration times of the tasks and average speeds of the drivers were logged automatically. The entire experiment was recorded with a digital video camera and a post-hoc analysis of the driving was used to evaluate the driver's performance. The experimental measures collected were the following:

1. Task completion time
2. Driving performance
3. NASA TLX workload test [84]
4. QUIS test [260]
5. Personal comments of the test subjects
6. Digital camera recording of the entire experiment

Experimental Procedure

Before the actual test all participants had time to adjust their sitting position and get acquainted with the interface controls (scrolling wheel, buttons, screen). Then subjects were asked to drive for approximately 5 minutes to get used to the driving simulator and the road conditions. The warm-up drive was followed by participants performing five tasks using the first interface. After a 15 minute break, the test subjects were asked to repeat the tasks with the second interface and, after an additional 15 minute break, with the third interface. No warm-up drive was performed before the second and the third set of tasks. To measure participants' perceived workload, after each set users were asked to fill in the NASA TLX workload test [84] and a slightly modified Questionnaire for User Interface Satisfaction (QUIS) [260]. For an overview on workload measurements please refer to section **??** in chapter 2. After the study participants were interviewed (unstructured) in order to collect their personal evaluation of the experiment.

Experimental Apparatus

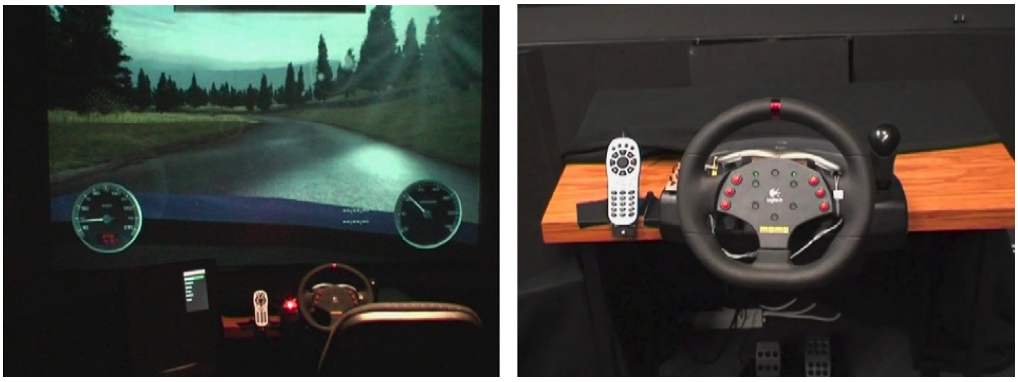
The experiment was conducted in a visualization room equipped with a large projection screen (2.4 m × 1.8 m) and a 7.1 surround sound system. The car simulation software RACER version 2.1[11] with the Swiss-Stroll[12] track was projected on the screen as can be seen in figure 7.16a). The simulator was controlled with the Logitech MOMO Racing steering wheel and automatic gear changing was applied (see figure 7.16b). The same type of car

[11] `http://www.racer.nl/`

[12] `http://www.racer-xtreme.com/prod/swiss-stroll/`

(Peugeot 307) was used throughout the entire experiment. As the experiment was performed in New Zealand the car was equipped for driving on the left-hand side of the road.

The Creative Sound Blaster X-Fi ExtremeMusic sound card and the Creative GigaWorks S750 speaker configuration system were used for sound reproduction. Loudspeakers were favoured over headphones, as blocking the auditory sense would have kept participants from parsing other co-occurring auditory events, such as the sound of the car engine, braking noises, and environmental sounds. The spatial sound generation was driven by the OpenAL sound library[13], which enabled access to all X-Fi hardware accelerated 3D sound features. The sound sources were the spoken items in the menu, recorded by a female native English speaker. The signal-to-noise ratio of the signals was approximately 50 dB.



(a) Screen, visual interface, steering wheel, and keyboard.

(b) Steering wheel and keyboard used for MSG task.

Figure 7.16: Car simulator used for the experiment.

Both acoustic menus were developed in the .NET programming environment. At each level in the menu, one to six sounds were generated and positioned at an equal distance around the user on a virtual circle (see figure 7.17). For example, if there were three items in the current menu, the spatial angle between the individual items was 120 degrees, while if there were six items in the menu the angle was 60 degrees. The centre of the virtual acoustic circle was positioned slightly to the back in order to put the listener

[13] `http://connect.creativelabs.com/openal/`

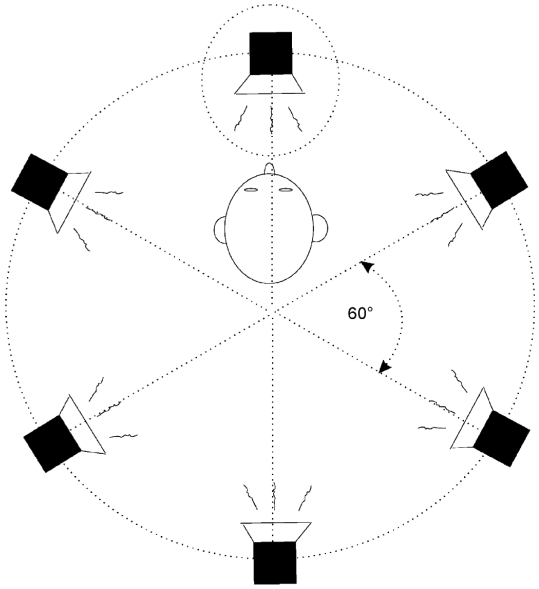


Figure 7.17: The layout of the virtual sound sources' positioning.

closer to the front items of the circle. The sound source positioned directly in front of the user (at 0 degrees azimuth) was always the one selected and therefore the loudest.

The visual menu as depicted in figure 7.13 was presented to the users on a 12 cm $\times$ 15 cm LCD screen with large white text on a black background, similar in style to the one used in Blasko & Feiner [261]. The current selection in the menu was highlighted with a light green bar. The application for the visual menu was developed using the .NET programming environment. All three menus could be controlled with the custom made navigation device depicted in figure 7.15. The navigation device consisted of a scrolling wheel and two buttons. The device was designed to be easy to operate while driving. The menu could be rotated in any direction with the use of the scroll wheel. In the visual menu condition this would cause the selection bar to move up or down and in the auditory conditions the virtual circle with sound sources to rotate clockwise or counter-clockwise around the user's head. The angle of the turn was always the angle between two neighbouring items in the menu so that one item was always selected. The left button confirmed the selected

item and loaded the items of the following level in the menu. The right button enabled a step back in the menu or cancelled the selected option. A small phone-like keyboard enabled text input when using the visual interface (see figure 7.13).

7.3.5 Results

As the data was numerical and normally distributed a one-way within subjects analysis of variance (ANOVA) with a fixed confidence level (p-value = .05) was used to analyse the data unless otherwise stated. Missing values/data points and/or outliers were removed from the analysis and hence the N may vary depending on the completeness of the data set.

Task Completion Time

The task completion time was measured between the initial command "Please start now." and the final notification "Task completed". The command was read to the users after the instruction on the individual task and the final notification was shown or played automatically when the task was completed. Figure 7.18 shows the average task completion times for the five tasks in the three different interface conditions.

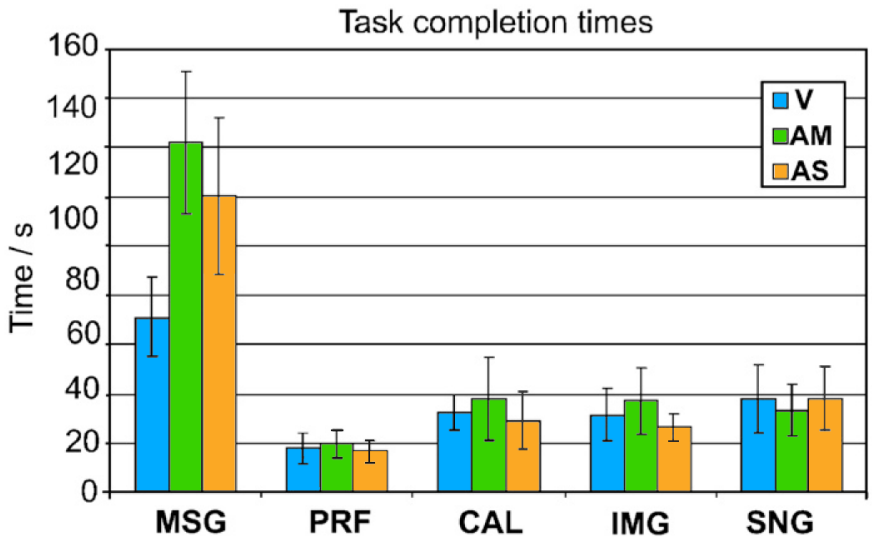


Figure 7.18: Mean task completion times for all tasks and conditions.

There was a significant difference in task completion times for the message composition task (MSG). The visual menu in combination with the mobile phone keyboard proved to be the fastest way to write a text message. An ANOVA for the MSG task gave the following result: $F_{MSG}(2, 51) = 8.52$, $p < .01$. A post-hoc Bonferroni test with a .05 limit on familywise error rate confirmed a significant difference between the visual (V) and auditory conditions (AM and AS), but no significant difference between the audio conditions AM and AS. The mean task completion times (in seconds) of the MSG tasks are presented in table 7.8.

Interface	$\mathbf{M}_{MSG}$	$\mathbf{SD}_{MSG}$
V	71.22	32.24
AM	120.50	63.45
AS	142.22	57.55

Table 7.8: Mean task completion times (M) and standard deviations (SD) for MSG task in seconds.

The reason for the visual interface being significantly faster may lie in the fact that most participants were already skilled in writing text messages with mobile phone keyboards. The audio interface for entering text messages turned out to be slow and quite cumbersome to operate in this particular environment. No significant difference between the individual interface conditions were found for the other four tasks:
$F_{PRF}(2, 51) = .36$, $p = .70$
$F_{CAL}(2, 50) = .55$, $p = .58$
$F_{IMG}(2, 51) = 1.21$, $p = .31$
$F_{SNG}(2, 50) = .21$, $p = .81$

Driving Performance

As the RACER simulation software did not support driving error logging, the driving performance was evaluated in a post-hoc analysis of video recordings of the subjects' performance. The users' driving during each individual task

was analysed and penalty points were assigned according to the following criteria:

- 1 penalty point: unsafe driving such as slight lateral deviation and slowing down unexpectedly and unnecessarily
- 2 penalty points: extreme lateral deviation and driving on the road shoulders
- 5 penalty points: causing an accident and crashing the car

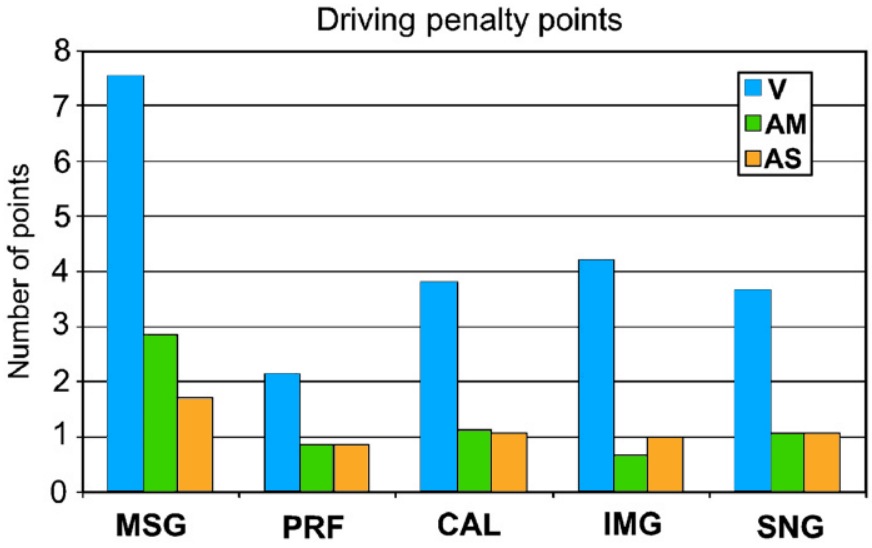


Figure 7.19: Mean driving penalty points for all tasks and conditions.

The penalty points for each driver were added up and the average penalty points for all users were calculated for each task (see figure 7.19). The number of penalty points was significantly higher in the case of the visual menu condition for these four tasks: MSG, CAL, IMG and SNG.
$F_{MSG}(2, 41) = 10.08$, $p < .01$
$F_{PRF}(2, 41) = 2.80$, $p = .07$
$F_{CAL}(2, 41) = 6.50$, $p < .01$
$F_{IMG}(2, 41) = 5.50$, $p < .01$
$F_{SNG}(2, 41) = 4.40$, $p < .05$

A post-hoc Bonferroni test with a .05 limit on familywise error rate confirmed a significant difference between the results of the visual and the audi-

tory interfaces, but no difference between the individual auditory interfaces (AS and AM). The mean values of the four tasks are presented in table 7.9. The average number of penalty points for the control group (test subjects who just drove the car) was .8. The mean values of the two auditory conditions as listed in table 7.9 and the control group show only a little difference - with the exception of the MSG task.

	M_{MSG}	SD_{MSG}	M_{CAL}	SD_{CAL}	M_{IMG}	SD_{IMG}	M_{SNG}	SD_{SNG}
V	7.53	5.11	3.08	3.32	4.2	5.22	3.67	4.3
AM	2.86	3.58	1.13	1.59	.67	.62	1.07	1.33
AS	1.17	1.32	1.07	1.68	1.00	1.66	1.07	1.43

Table 7.9: Mean driving penalty points (M) and standard deviations (SD) for the tasks: MSG – composing and sending the message; CAL– making a call to a specific person; IMG – deleting a specific image; SNG – playing a specific song.

	MSG (%)	PRF (%)	CAL (%)	IMG (%)	SNG (%)
AM to V	71	33	78	50	66
AS to V	55	57	64	77	59

Table 7.10: The relative improvement of the driving performance comparing the auditory conditions AM and AS condition to the visual V condition.

However, due to low number of participants in the control condition, a statistical evaluation cannot be conducted. The driving improvement I in the auditory conditions can be corroborated by calculating the relative change of the driving penalty points, comparing the AM or AS condition to V condition. The relative change I_{AM} and I_{AS} per user per task can be defined as

$$I_{AM} = \frac{D_{AM} - D_V}{D_V} \quad and \quad I_{AS} = \frac{D_{AS} - D_V}{D_V},$$

where D_{AM} and D_{AS} are the number of penalty points when using each of the auditory interfaces and D_V is the number of penalty points when using

the visual interface. The mean values of driving improvements of all users are presented in table 7.10: The average improvement in driving performance over all tasks in the AM condition compared to the V condition is 62 percent and 60 percent improvement in performance comparing the AS condition to the V condition.

Workload

Workload was measured with the NASA TLX (for Windows), a multi-dimensional rating procedure that derives an overall workload score based on a weighted average of ratings on six sub-scales:

1. **Mental demand**: How much mental and perceptual activity was required (thinking, deciding, calculating, remembering, etc.)?
2. **Physical demand**: How much physical activity was required (pushing, pulling, turning, controlling, etc.)?
3. **Temporal demand**: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?
4. **Performance**: How successful do you think you were in accomplishing the goals of the tasks set by the experimenter?
5. **Effort level**: How hard did you have to work (mentally and physically) to accomplish your level of performance?
6. **Frustration level**: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content and relaxed did you feel during the task?

Figure 7.20 shows the final overall workload scores with standard deviations for all three interfaces.
An analysis of variance showed a significant difference in the overall workload between the three conditions with $F(2, 321) = 15.39$, $p < .01$. A post-hoc Bonferroni test with a .05 limit on familywise error rate confirmed that the workload reported in the visual condition (V) was significantly higher than the workload generated in the AM condition ($p < .01$) and in the AS condition ($p < .01$). Only a near significant difference between the two auditory conditions could be found ($p = .053$).

Further examination of the results of the individual sub-scales of the TLX workload test revealed some interesting results. There is a significant difference between the conditions in the following four sub-scales:

- **Physical demand**: $F(2, 51) = 4.09$, $p < .05$;
- **Temporal demand**: $F(2, 51) = 4.65$, $p < .05$;
- **Performance**: $F(2, 51) = 4.24$, $p < .05$;
- **Frustration**: $F(2, 51) = 3.19$, $p < .05$.

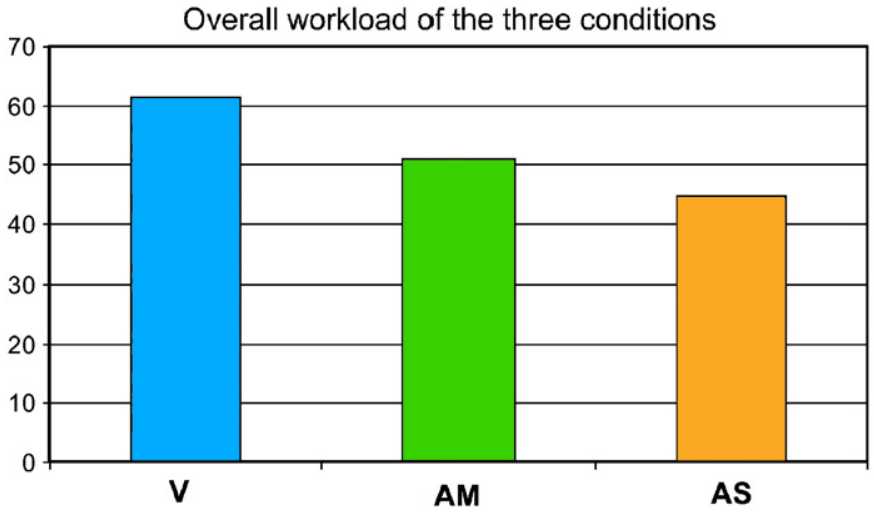


Figure 7.20: Mean values and standard deviation of the final TLX workload test (V - visual menu; AM - auditory menu with multiple; AS - auditory menu with a single sound).

The post-hoc analysis showed that, in all four cases, the visual (V) condition differed significantly from the two auditory conditions; no significant difference between the auditory conditions AM and AS could be confirmed.

User Interaction Satisfaction

Participants' subjective ratings were measured with a reduced version of the Questionnaire for User Interaction Satisfaction (QUIS) [260]. The QUIS assesses users' subjective satisfaction with specific aspects of a human–computer interface. As the data were normally distributed and the 9-point Likert scales

used in the questionnaire delivered an interval-level measurement the average scores of all three interfaces were compared through an ANOVA and a post-hoc Bonferroni test.

The auditory interfaces performed significantly better than the visual interface for the following items. Mean values (M) and standard deviations (SD) for all items are listed in table 7.11:

1. wonderful – terrible: ($F(2, 51) = 9.40$, $p < .01$)
2. easy – difficult: ($F(2, 51) = 14.17$, $p < .01$)
3. satisfying – frustrating: ($F(2, 51) = 7.41$, $p < .01$)
4. adequate – inadequate: ($F(2, 51) = 11.81$, $p < .01$)

The scores were not significantly different when the users were asked about the following word pairings:

5. stimulating – dull: ($F(2, 51) = 3.14$, $p = .052$)
6. flexible – rigid: ($F(2, 51) = 2.50$, $p = .09$)

No significant difference were found between the conditions concerning the participants' ability to:

7. learn to operate the system: ($F(2, 51) = 1.08$, $p = .35$)
8. explore new features by trial and error: ($F(2, 51) = 2.15$, $p = .13$);
9. remember names and use commands: ($F(2, 51) = 1.53$, $p = .23$).

	Wonderful – Terrible		**Easy – Difficult**	
	M_1	SD_1	M_2	SD_2
V	3.06	2.07	2.22	2.42
AM	5.11	2.22	5.11	2.65
AS	6	1.97	6.39	2.43

Satisfying – Frustrating			**Adequate – Inadequate**	
	M_3	SD_3	M_4	SD_4
V	3.5	2.3	2.83	2.31
AM	4.89	2.02	5.39	2.48
AS	6	2.3	6.33	1.88
Stimulating – Dull			**Flexible – Rigid**	
	M_5	SD_5	M_6	SD_6
V	4.33	2.27	4.44	1.76
AM	5.94	1.82	5.17	2.23
AS	5.61	1.98	6	2.25
Easy – Difficult to learn			**Easy – Difficult to explore**	
	M_7	SD_7	M_8	SD_8
V	7.33	2.00	7.17	2.06
AM	6.39	2.23	5.67	2.45
AS	7.00	1.61	6.39	1.98

Easy – Difficult to remember		
	M_9	$S.D._9$
V	6.83	1.89
AM	5.67	2.17
AS	6.56	2.20

Table 7.11: Mean values and Standard Deviations for pairings from the Questionnaire for User Interaction Satisfaction.

The results show that participants had a high overall satisfaction with all three interfaces. They found the auditory interfaces to be rather easy to use, satisfying, and adequate. On the other hand, users did not find them

significantly more stimulating or flexible than the visual interface. As regards the learning required to use the interfaces, users reported all interfaces to be equally difficult to learn; features seem equally difficult to explore and names and commands equally difficult to remember.

User Comments

After each experiment participants were interviewed about their experience. In this section the most frequent and most insightful positive and negative comments are listed.

The visual interface (V):
Positive

- "The interface was very simple to use."
- "There was better information on the current position in the menu."
- "It was faster than the auditory interfaces."

Negative

- "It demanded full attention for operation."
- "The users had to wait for an "easy" segment of the road to complete the tasks."
- "It was very distracting and dangerous."

The auditory interface with multiple sounds (AM):
Positive

- "It was very easy to drive and complete the tasks simultaneously."
- "The drivers could keep their hands on the wheel."
- "It was very useful, especially for short tasks."

Negative

- "It was hard to listen to, especially when the engine noise was loud."
- "There were too many different sounds at the same time."
- "There was no overview of the entire menu structure."

The auditory interface with one sound (AS):
Positive

- "It was less distracting than the visual interface."
- "It was easy to understand and adapt to."
- "It was less confusing than the interface with more sounds."

Negative

- "Writing a message with the acoustic menus was too complicated and took too long."
- "There was no good feedback on the entered words."
- "There was no information on the current position in the menu and the users sometimes had to scroll through all the items."

7.3.6 Discussion

The main goal of this study was to evaluate in a semi-realistic "mobile" environment, how task completion time, primary task performance, workload, and user interaction satisfaction are influenced by different interface types, namely two eyes-free and one traditional head-down visual interface (V). The two acoustic interfaces were modeled on the interface described in 7.1 and differed only in the number of simultaneously playing sources: just one source (AS) or up to six sources (AM). All three interfaces consisted of the same hierarchical menu structure simulating a common mobile phone interface structure and were controlled with the same custom-made interaction device.

No significant differences in task completion times were found, except for the MSG task, whereby users had to enter and send a text message to a specific person. The longer task completion time in this case is a consequence of the use of different and unequally efficient interaction devices: a mobile phone keyboard in the visual condition and an auditory menu for text entry in the auditory conditions. Although speech recognition would have been a more elegant and effective way of entering text, to maintain comparability an interaction procedure similar to that of manual text input on a keyboard was designed for the auditory interfaces.

Since entirely new acoustic interfaces were compared to a well-known and widely used visual interface, the similar task completion times across the conditions are encouraging for the use of auditory interfaces in vehicles. The initial hypothesis concerning significant improvement of the driving performance in the two auditory conditions were justified. Participants drove more safely when operating the auditory interfaces; the average number of penalty points dropped by 60 percent in the audio conditions when compared to the visual interface condition.

A comparison of the driving performance of the control group, in which participants did not perform secondary tasks while driving, to the participants who performed tasks with the two auditory interfaces shows no degradation, although the driving speed of the control group was, on average, higher.

Participants found performing the tasks with the visual menu rather difficult, dangerous, and unpleasant. In some cases, participants slowed down or even stopped the car to perform the task safely, which caused large variations in the driving speed. In the case of the two auditory interfaces, these variations were negligible.

Participants reported a significant difference in the perceived workload between the three conditions. In general, the results of the NASA TLX indicate that participants felt less physical and temporal demand when interacting with the auditory interfaces. They felt a high level of satisfaction and were confident about their performance. The use of the auditory interfaces made them feel more secure and less stressed than the use of the visual interface. This is contrary to our initial expectation that participants would find the new auditory interfaces harder to use and more difficult to adapt to.

The results of the questionnaire on user interaction satisfaction showed high satisfaction among participants when using the auditory interfaces. Subjects reported that the auditory interfaces were easier to use, more satisfying, and more adequate than the visual interface. Most test subjects commented on the importance of learning effects in the experiment, especially with the auditory interfaces. The visual interface was more effective and easier to use in the beginning, but the auditory interfaces became as effective after a few uses. The users reported that the auditory interfaces could be quite confusing

when performing longer tasks that required a lot of movement through the hierarchical menu structures. Participants reported having difficulty with orientation within the menu structure, which was not the case in the visual interface. The latter was confirmed with the last three results of the post-study questionnaire where participants reported the visual interface to be easier to learn and easier for exploring new features.

In the AM interface, all items of the current menu level were played simultaneously in an attempt to present as much information as possible at a given time, while in the AS condition menu items were played one at a time. Participants reported the AS condition to be more effective, since it made it easier to concentrate on the driving while most of the sounds in the AM condition were perceived as background noise and not as additional overview information on the contents of the menu.

While faster task completion times for the AM condition were expected, the inability to extract helpful information from the additional sound sources is reflected in only minor differences in task completion times between the auditory conditions. Apparently, the cognitive load[14] generated by attending to up to six sound sources was so high, that it competed for the attention required to steer the vehicle. Rationally, participants neglected the offered *plus* of information in the AM condition in favour of driver safety and focussed only on the information necessary to complete the secondary task. Perhaps up to three sounds played at once could have the advantage of enabling a larger information flow without generating too much cognitive load. Further investigations into ways to increase the information density in auditory interfaces while not affecting the driver performance are required.

The awareness of the current position within the menu hierarchy proved to be the biggest disadvantage of the auditory interface compared to the visual interface. Therefore, additional acoustic cues and frequent feedback messages should be added to the interface. The ineffective text input mechanism made the acoustic interface much slower for longer tasks where participants had to compose a message or enter some commands manually. In the future, this should be replaced by a voice recognition system or predictive text editing.

[14] Please refer to section 2.2 for a summary of the Cognitive Load Theory (CLT).

It would be interesting to test realistic road conditions. The car simulator used in this study consisted of a country road without other cars or more dynamic obstacles on the road. A real car driven through a city centre under various traffic conditions would demand an even higher degree of user concentration and would allow for the gain of interesting insights into the efficiency of the next generation of auditory interfaces.

7.3.7 Conclusion & Design Recommendations

While the first two sections of this chapter addressed **RQ 2, 2.1** and **5** the study described in this section mostly addressed **RQ 2**: What are viable non-visual multimodal interaction techniques? and **RQ 3**: What is a good way to help users obtain an overview of available items and options? – in a special case of mobile UI, i.e. an in-vehicle user interface. Besides addressing these superordinate research questions, one of the key functions of this study was to explore how eyes-free interfaces, compared to a traditional head-down visual interface, impact primary task performance.

The results of the study suggest that the auditory interfaces tested are at least as suitable for use with in-vehicle information systems as visual interfaces. Auditory interfaces could significantly contribute to driver safety since they do not compete for the same sensory resources as visual interfaces.

An auditory interface with verbal output proved to be very effective for shorter tasks such as changing the settings, selecting songs, or making a call to someone. However, good feedback on the current position of the user in the menu should be given in order to avoid confusion and the need to reset the system to reorientate. The background music with a changing central pitch turned out to be a good solution to help the user identify the individual sub-menus, but it should be upgraded with some spoken feedback options. For example, the option "current location" could inform a user of their current position and even available commands.

The auditory text input system proved to be too slow and therefore inappropriate for composing messages or performing longer tasks that demand the input of text. An effective voice recognition system would be a better solution, but problems could emerge from the noisy in-vehicle environment

consisting of the operating noise, but also passengers, and entertainment devices. The interaction device built for this study was very effective and turned out to be very appropriate and easy to use while driving a car. The users found it safe to use, since they could have both hands on the steering wheel at all times.

Finally, it is important to emphasize that multiple simultaneous sounds, in the form that they were implemented in this study (i.e. the AS condition), might not be the best UI design choice for performing secondary tasks while being engaged in a challenging primary task such as driving. The anticipated benefits of creating an overview of available items by playing sources simultaneously could not be verified. On the contrary, the high cognitive workload created is likely to have a detrimental effect on the primary task performance and has a high potential to annoy and/or stress users.

Chapter VIII

Foogue: An Eyes-Free UI Design Concept

In this chapter, *Foogue* is presented, a design concept for an eyes-free interface that utilises spatial audio and gesture input. *Foogue* was designed with the psychoacoustic and cognitive factors summarized in chapter 2 in mind and most of the findings stated in previous chapters 3, 4, 5, 6, and 7 are incorporated into the design of *Foogue*. It is important to point out, however, that *Foogue* is a design study and a proof of concept, not a finished interface or product. The work presented in this chapter has neither been implemented nor evaluated. It integrates many of the results gained through studies described in previous chapters, but it lays no claim to be comprehensive.

8.1 Introduction

Many Smartphones come with functionality that is comparable to that of mobile computers. So far their interfaces have not deviated far from the WIMP paradigm and graphical user interfaces (GUIs) are still predominant. The user can switch between 'views' or 'screens', and icons and hierarchically structured menus are widely used. The mouse has been replaced by a stylus, touch or multitouch interaction. Although GUIs are highly efficient in desktop computing and have a long history of research and optimization, in many countries using a smartphone while driving is banned. This is due to three drawbacks visual interfaces have in mobile situations: Firstly, as a result of the limited screen size only a little information can be displayed. Secondly, to retrieve the information the user has to hold the device up close and focus on the screen. As most mobile situations, like driving a car or navigating through an urban environment, require visual attention, the consequences of distractions caused by looking at the screen (and not focusing on the task at

hand) can be severe. Thirdly, most feedback is presented visually. Entering letters, selecting icons or scrolling require the user to continuously look at the screen while interacting with the device. This either forces the user to disrupt the primary task (e.g. stop walking / driving) and hence to turn the mobile situation into a stationary one, or it diverts visual attention away from the primary task and results in the same conflict described above.

Using headsets or hands-free kits for phone calls partly solves the haptic distraction caused by holding the phone to the ear. Miniaturizing icons or offering several selectable 'screens' makes more efficient use of the limited screen size. Adding additional buttons or using regions on the screen to access frequently used functions reduces the 'eyes on screen' time for these few functions. However, neither of these solutions overcomes the distraction caused by pursuing two or more competing visual tasks at the same time.

This issue is addressed with *Foogue*, a 3D audio interface that supports menu navigation, item selection, and 'window' management via haptic interaction. While audio has been widely used for alarms, notification, and feedback, *Foogue* offers a spectrum of functionality that is comparable to common visual interfaces. By employing audio and haptic interaction, *Foogue* avoids sensory conflicts with visual tasks. *Foogue* also enables visually impaired users to fully access mobiles phones and it is transferable to other mobile or stationary devices, such as tablet computers or laptops.

8.2 Related Work

Related to the concepts presented in this section are findings from a range of different disciplines and research approaches:

- A broad overview of relevant psychoacoustic topics, concepts, and theoretical foundations is given in chapter 2
- The efficiency and user friendliness of gestural or tactile input in combination with auditory interfaces has been thoroughly explored in chapter 7. Prior related work on this subject includes [16, 167, 262, 15, 6, 177, 241, 242, 240]
- Overview and monitoring techniques for auditory interfaces have been explored in the context of an AR application in chapter 4. Relevant

prior work in this field is [209, 152, 153, 151, 203, 185, 204, 154]
- Interface metaphors are discussed in section 2.6.4 of chapter 2. Examples of prototypes deploying a circular layout are [168, 170, 165, 14, 166, 167, 169]
- General issues impacting the pleasantness and appropriateness of using spatial sound in auditory displays were addressed in chapters 5 and 6.

The Nomadic Radio by Sawhney & Schmandt [14] was especially influential to the design of *Foogue*. The Nomadic Radio is a mobile, shoulder-worn speaker and microphone system which allows the wearer to manage voice and text-based messages, including voicemail, email, calendar entries, news, traffic, and weather updates. The user interacts with the system by either voice command or tactile input. The Nomadic Radio is based on a holistic approach and is strongly oriented towards content accessibility, process and task convergence and, moreover, it has an activity and context awareness that many later systems lack.

Foogue also incorporates many features proposed by the author and colleagues in [6] and summarized in section 7.2. Here, the user can either use gestures with the mobile phone or press keys to interact with an auditory display consisting of multiple sound streams positioned on a circle around the user's head. In a manner similar to the one described in this system, *Foogue* strives to achieve a balance between straightforward, fast, and rich access to information and the unobtrusiveness required to keep the user "undisrupted" and free from information overload. Part of this approach is an effective multitasking and attention management that allows the user to focus part of their attention on a specific task or stream of information while other, less important streams can still be monitored. In section 7.2 and chapter 3 the notion of distance has been explored as a means to manage attention and as a metaphor to convey attributes such as importance or similarity. This concept is picked up and further elaborated in *Foogue*. Also, the lessons learned from prior interface designs (as studied in sections 7.2 and 7.3) are applied to the current project. These include, but are not limited to:

- For simple operations 2D gestures performed on a touchscreen are preferred over 3D gestures performed with the device. For complex oper-

ations combinations of 2D and 3D gestures are viable.

- Consistent feedback for successful and unsuccessful operations is essential.
- Gesture analogies from other technical domains or devices are exploitable.
- Discreet gestures are preferred over very expressive gestures.
- Gesture inversions for do-undo-commands are recommended.
- Forming a mental model of a system is easier and more accurate when it is not hindered by learning of a new interaction technique at the same time, i.e. learning to operate a complex system using buttons on a keyboard is easier than learning how to use it by gesture interaction.

A third inspiration for *Foogue* is Shoogle [15] because of its playful approach and subsequent employment of an interface metaphor. In Shoogle, the mobile phone or PDA becomes a container, a *box*, that can be shaken to reveal its contents. Elements, such as messages, become *balls* that move inside the container. Through acoustic and vibrotactile feedback, the authentic behaviour of objects is simulated and, from everyday experience, users can deduce the relative quantity of items contained when shaking the device. The metaphor chosen is so simple yet suitable that once the metaphor is understood, anybody who ever held a box of matches in their hand should be able to use Shoogle successfully.

In the following, the core concepts are explained, rather than the particulars and details of the proposed solution. Although most user input is referred to in terms of gesture, many gestures can be substituted or replaced by keyboard interaction to enable operability on devices without a touchscreen or, to support user preferences.

8.3 Interface Design

8.3.1 Modes

Foogue features two modes. Modes are distinct settings within a human-computer interface, in which the same user input leads to different results in different modes. A well known example is the Caps lock button on a keyboard, that, when pressed, renders typed letters to uppercase by default.

Although modes should generally be avoided due to the error proneness they entail [263], *Foogue* has one mode to support the predominantly active phase when interacting with the menu (*Menu Mode*) and a second to support the predominantly passive phase of listening to selected items (*Listening Mode*). The user switches between modes by performing the *switch mode* gesture depicted in 8.7. Changing between modes is confirmed by sound feedback. If the user switches from *Listening Mode* to *Menu Mode*, all sound sources are paused and the menu is displayed according to its last state.

Menu Mode

In this mode the user navigates, selects or manipulates items from a hierarchical menu. It is comparable to using the 'Explorer' in Microsoft Windows or the 'Finder' in OS X. As depicted in figure 8.1, in *Menu Mode* all items are spatialised and arranged in a 120 to 180 degree arc in front of the user. Keeping items within that range prevents front-back confusion and keeps pointing gestures within the normal movement radius of the wrist joint.

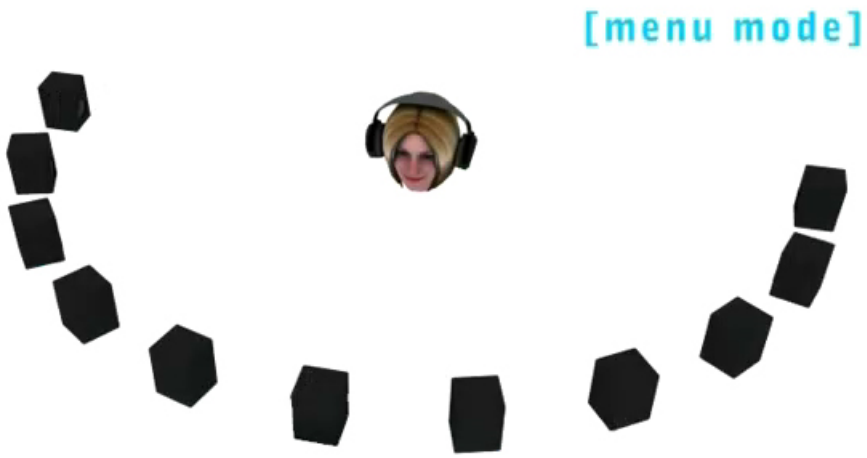


Figure 8.1: *Menu Mode*: A list of files available to a user.

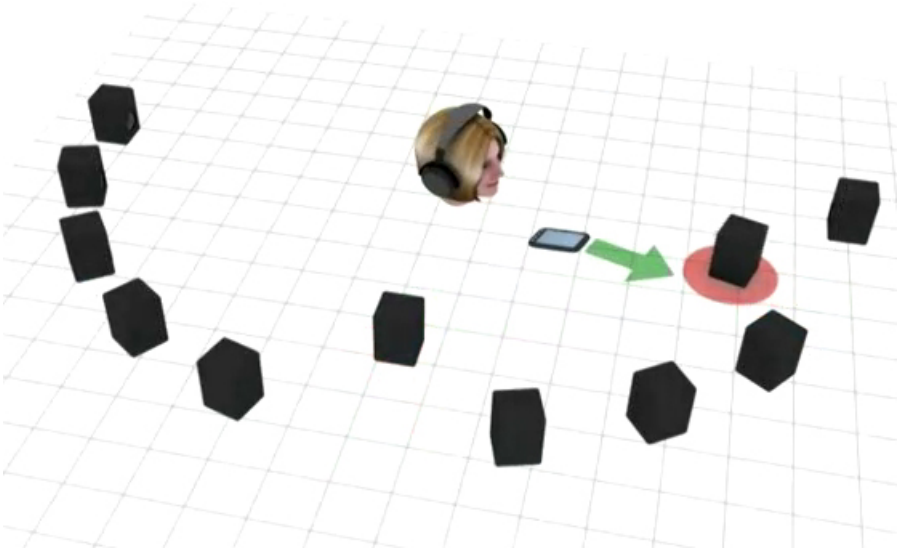


Figure 8.2: *Menu Mode*: A user selecting two items from a list.

Items are displayed in sequence and ordered alphabetically, but can be alternatively ordered by file type, size or recent use status. Items' positions are fixed to use spatial memory and thus, allow users to 'jump' to a specific item without the need to scan through all displayed items. When pointing the device at an item, the item identifies itself by speaking its (file-)name (see section 8.3.4 for examples). *Foogue* supports single and multiple item selection as well as the selection a range of items. For single file selection the user points the phone at the item and performs the *open* gesture. For multiple file selection – from either one or multiple folders – the user points at items and moves them individually to the *buffer*, a zone around the user (as shown in figure 8.2). Whole folders can also be pulled into the *buffer*. Figure 8.5 depicts the *push* and *pull* gestures. When in *Menu Mode*, pulling items closer with the *pull* gestures moves them to the buffer. Pushing them away with the *push* gestures removes them from the buffer. *Foogue* incorporates gesture reversibility, i.e. the reversed gesture undoes the prior action. The same logic applies throughout the system: if moving a source closer to the user adds it to the *buffer*, moving it away removes it from the *buffer*.

For selecting a range of files the user points at the start item, performs the *select range* gesture and moves the device to the end item. If the *open*

gesture is performed either on a single file or the buffer, the appropriate *player* is instantiated in *Listening Mode* and the file/s is/are played when this mode is entered.

Players are one of the core concepts of *Foogue*: Analogous to *Windows*, they give access to and display content. By instantiating a *Player*, that is performing the *open* gesture on any item, *Foogue* chooses the appropriate application automatically. This can be a text-to-speech engine reading the content of a file, a media player, or a phone or data connection.

Listening mode

In this mode the users mainly listen to what they have previously selected. Nevertheless, they can interact with *players* and rearrange their positions. Unlike in *Menu Mode*, *players* can be positioned anywhere on a 360 degree circle around the user. Figure 8.3 illustrates the concept.

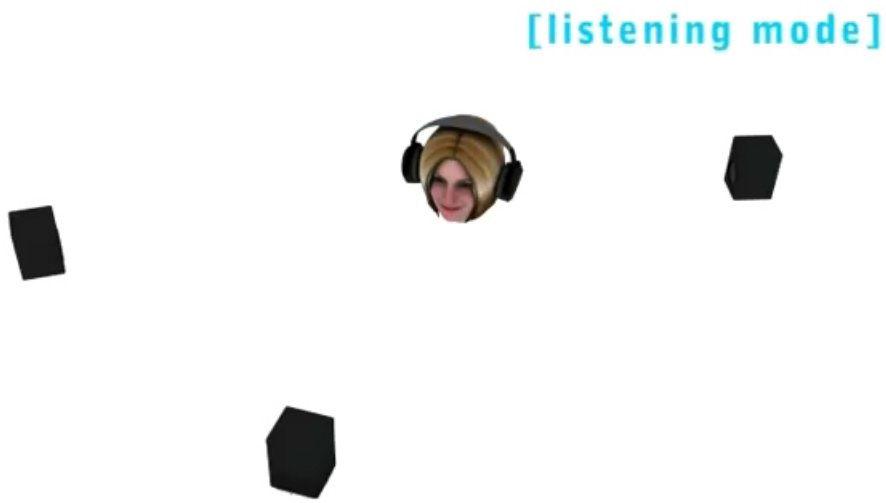


Figure 8.3: *Listening Mode*: Players available to a user.

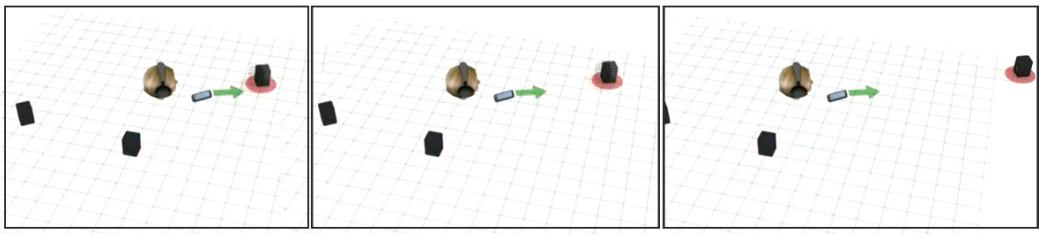


Figure 8.4: A user selecting and repositioning players in *Listening mode*.

Players can be clustered – for example all (sporadic) notifiers on the left and continuous playback (like music or podcasts) on the right side. Not only can *players* be rearranged in terms of direction but also in terms of distance. *Foogue* uses distance as a metaphor for minimizing/maximizing or, in other words, to focus/defocus attention. If the user wants to stay aware of a *player*, such as one displaying notifications, it can be moved farther away. Figure 8.4 illustrates how a user can employ the *push* gesture to move a *player*. This way distraction from that player is minimized while the user is still able to maintain a peripheral awareness of the *player*. If the user wants to focus entirely on one *player*, as during a phone call, the user can perform the *open* gesture on the *player*. To keep the set of gestures small and recognizable, while in *Listening Mode*, the *open* gesture 'activates' a *player*, i.e. the selected *player* is focused upon while all other *players* are paused (when routed through a text-to-speech engine or media player) or muted (when live, like in a conference call). Activating a *player* has three effects: All other players are paused/muted, the stream from the *player* is played in stereo (if available), and the context or 'right-click' menu is displayed on a 120 - 180 degree arc just like in *Menu Mode*. As items in the context menu are silent unless they are pointed at, this keeps the context menu available without causing distraction.

8.3.2 Interface Metaphors

Lakoff & Johnson [12] describe how thinking in ontological metaphors enables humans to refer to otherwise abstract concepts (see 2.6.4 for an introduction to interface metaphors.). Thinking of data in terms of an object that can be

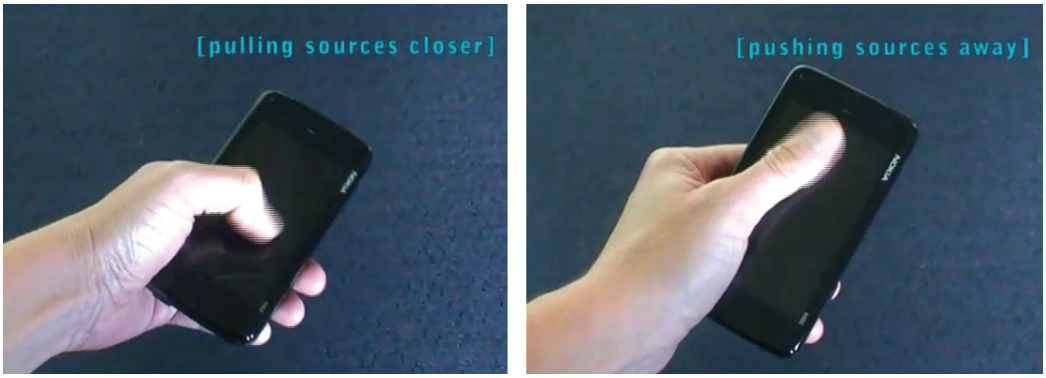


Figure 8.5: Gestures to bring *players* closer or push them away.

moved, copied, or named is a simple example of an entity metaphor. *Foogue* uses three ontological metaphors: Containers, entities, and substances. An example of a container is the *folder*, which contains either other *folders* or *files*. In *Menu Mode files* are thought of as entities. Users can navigate through a structure of *folders* and select, copy or move *files*. But in *Listening Mode*, when a file is 'played', it changes its nature and becomes a 'substance' (water). It is routed through a *player*, it can be 'diverted' (moved), its 'flow' can be 'disrupted' (paused) and so forth.

The difference in purpose of both modes is reflected in the metaphors applied. *Menu Mode* supports the perception of data in terms of a solid structure of containers and objects. But the actual playback in *Listening Mode* refers more to the temporal, fluid nature of sound. Besides that, so called 'orientational' metaphors are used in *Foogue*, which refer to spatial orientation such as *front/back* or *up/down*. The *close/far* or *central/peripheral* metaphor was described above in the context of focusing attention via positioning items closer or farther away. Pushing *players* away implies a 'downgrading' in terms of the amount of attention paid to them. By pulling a *player* closer it gets full attention and accordingly all other *players* are paused. The *up is more* metaphor is used when volume is regulated: performing the *more* gesture depicted in 8.6 on a player, will increase the volume while the inversed gesture will decrease the volume. The *more* gesture is modeled on the *pull* gesture described in section 7.2.

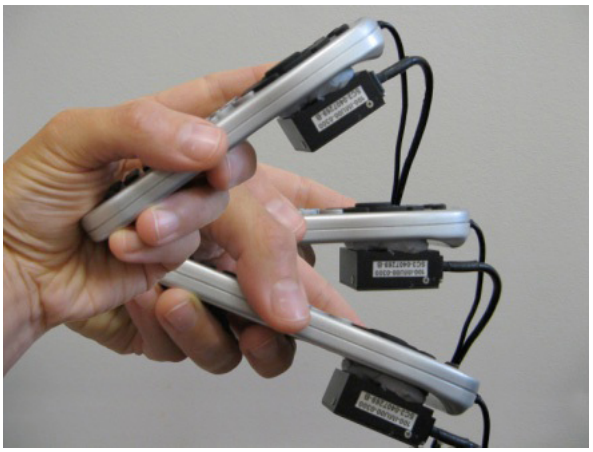


Figure 8.6: *More* gesture: A user tilts the phone up to increase the volume of a player and tilts the phone down to decrease the volume of a player in *Listening Mode.*

8.3.3 User Input

A gesture language is proposed that is built from a limited number of simple and easy to differentiate gesture elements. This language is a combination of keyboard and 2D and 3D interaction techniques, such as the *point*, *tilt*, *move*, *rotate*, *drag & drop* gestures similar to the user-designed gestures created in the task-driven explorative experiment described in section 7.1 of chapter 7. When elaborating gesture design definitions, special emphasis has to be put on designing gestures that are intuitive and easily discoverable, but that also take into account the limited movement range of mobile users. While a few central gestures have been designed, the remaining corpus remains to be specified in a future developmental effort.

The *open* gesture mentioned above has an essential similarity with the 'double click' performed with a mouse and is therefore a 'double tap' on the touch screen. Given the results from the user study summarized in section 7.1 users are able to transfer concepts between different interfaces. Therefore, it makes sense to exploit the users' already existing association between a 'double click/tap' and the resulting action of opening a file and transfering it to the rather new domain of auditory interfaces. The *change mode* gesture depicted in figure 8.7 is a 90 degree rotation of the phone to the side. This

Figure 8.7: *Change Mode* gesture: A user rolls the phone 90 degrees and changes from *Menu Mode* to *Listening Mode.*

is, in contrast to the rather universal *open* gesture, a unique gesture that only changes between modes. When in *Menu Mode* the gesture changes the selected mode to *Listening Mode* while preserving *Menu Mode's* state and vice versa. To prevent accidental user input the *lock/unlock* 2D gesture shown in figure 8.8 can be performed on the touchscreen. The 'Z' shaped pattern locks or unlocks the screen. This gesture is comparable to the keyboard or screen lock feature on most mobile phones.

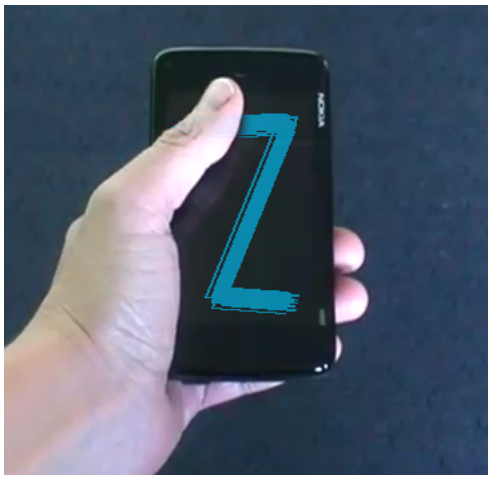


Figure 8.8: *Lock/unlock* gesture: A 'Z' touch gesture on the screen locks or unlocks the device.

8.3.4 Next Steps

Besides already-existing components of *Foogue*, two prototypes were developed to research selection procedures from uniformly and non-uniformly distributed lists. When trying to find a given key item in an address book or playlist, which is essentially a sorted list, users often follow an interpolation search pattern[1]: after the initial guess users compare the key found with the key they were looking for and correct the position they look at in the next step until they find the specific key they were looking for. For example: If a user wants to call a contact named "Jones" from an alphabetically ordered contact list, their initial estimation may be that names starting with 'J' are about one third of the way through the list. They may then point at an item in that region and find the name "Miller". They now know that they have to correct slightly towards the beginning of the list and may find the name "Jaspers" in the next step. In a third step they will correct towards "Miller" as "Jones" is listed between "Jaspers" and "Miller", and so forth.

On average the interpolation search makes $\log(\log(n))$ comparisons if the elements are uniformly distributed, where n is the number of elements to be searched [264]. However, names in a contact list will not be uniformly distributed as some names are much more common than others (like Smith or Johnson)[2] and some initial letters (like X or Y) are less common than others depending on the language of interface. If the distribution is non-uniform and the list very long, users may choose an interpolation-sequential search, that is, begin with an interpolation of the approximate location of the key item, and then proceed with a linear search until they find the actual location.

Two prototypes were built in order to study the search patterns of users in both uniformly and non-uniformly distributed, aurally presented, lists. While in a first exploratory study the search patterns would be of primary interest, in later studies the impact of the interaction technique and the

[1] By comparison, binary search always chooses the middle of the remaining search space, discarding one half or the other of the available search space, again depending on the comparison between the key found at the estimated position and the key sought.

[2] `http://en.wikisource.org/wiki/1990_Census_Name_Files_dist.all.last_(1-100)`

impact of the spatial arrangement of the sources ought to be investigated.

The prototype shown in 8.10 was built to evaluate human search patterns for an acoustically presented uniformly distributed list of numbers between 1 and 500. The prototype is written in Java and using the FreeTTS[3] speech synthesizer to generate spoken words from the number set. While the functionality of the prototype allows for a spatial arrangement of the generated sound sources (by passing it through OpenAL[4]), for an initial study all sources are aligned horizontally and their distance from each other corresponds with the item-to-window size ratio seen in figure 8.10; with a window size of 500 pixels and a total of 500 numbers, moving the cursor from one pixel to another corresponds with selecting the neighbouring sound source. The prototype supports item selection by using either gestures or a computer mouse. Therefore, not only can the search pattern be studied but also the impact of the interaction technique on the user behaviour.

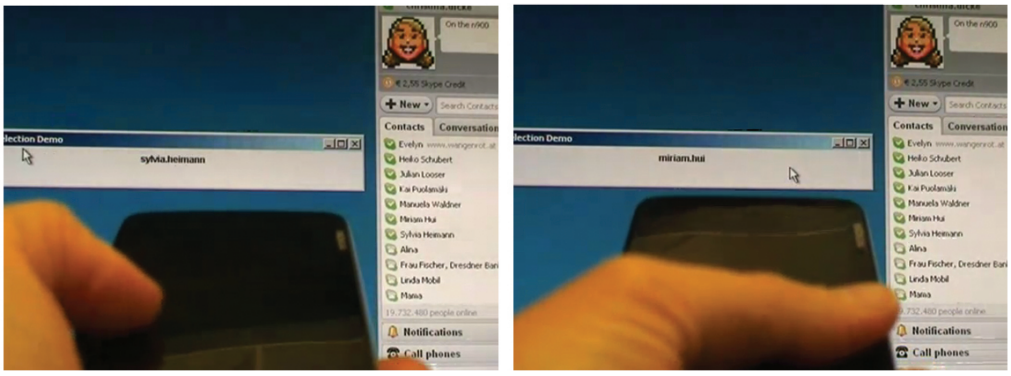


Figure 8.9: Skype interface demo: A user selecting a contact from their Skype online contacts list.

The prototype shown in 8.9 generates a non-uniformly distributed list of names from the contact list of a Skype[5] user. The contacts' handles are acquired as strings via the Skype4Java Java bindings for Skype, passed to FreeTTS and are then positionable in 3D space via OpenAL. In both cases,

[3] `http://freetts.sourceforge.net/docs/index.php`

[4] `http://connect.creativelabs.com/openal/default.aspx`

[5] `http://www.skype.com/`

the playback of the item's value or name is interrupted when the user moves the pointer to another item. In this way the process of 'scanning' or skipping through the list is accelerated.

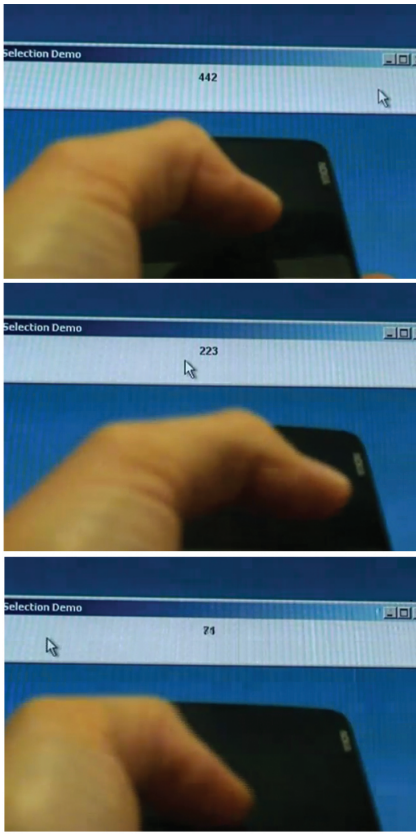


Figure 8.10: Item selection demo: A user selecting one item out of 500.

8.3.5 Technical Feasibility

Using inbuilt gyroscopes, digital compasses or cameras (visual feature tracking) the smartphone's position in space can be determined. Devices like the iPhone already supports the 3D sound library OpenAL and the Android Platform will support the 3D sound library OpenSL ES[6] from version 2.3 "Gingerbread" onwards.

[6] `http://www.khronos.org/opensles/`

The prototypes described above are implemented on a PC using an Intersense InertiaCube3[7] featuring an accelerometer, gyroscope, and magnetometer for 3-DOF user input, FreeTTS for text-to-speech generation, and OpenAL for creating the 3D sound space.

[7] http://www.intersense.com/pages/18/11/

Chapter IX

Conclusions

9.1 Summary of the Thesis

This thesis explored the possibility of using mobile eyes-free interfaces in situations where traditional visual interfaces are not appropriate. In an introductory chapter 1 the motivation for the work was discussed and several research questions were raised. Starting from the assumption that everyday perception is formed in response to conceptual metaphors and that often abstract concepts are thought of in terms of objects [12], the success of the WIMP paradigm was discussed. It was argued that the WIMP paradigm addresses and solves many key issues in human-computer interaction, such as representing data as entities, supporting navigation in hierarchies, and object selection and manipulation. While the original implementation of the paradigm usually featured graphical user interfaces, a fixed screen, and a mouse, interfaces for mobile devices implemented the WIMP paradigm without acknowledging the different usage scenarios and hence the different usability requirements.

The research objective of this thesis is to provide a critical evaluation of alternative presentation and interaction methods taking into account both the limitations and possibilities of mobile devices as opposed to stationary PCs. The following research questions were raised and addressed in the different chapters of this thesis:

RQ 1 What are the advantages and disadvantages of using sound?

RQ 1.1 How can spatial sound be utilised?

RQ 1.2 What are the advantages and disadvantages of using spatial

sound compared to stereophonic or monophonic sound?

RQ 1.3 How can acoustic distance perception be used as an aspect of interface design?

RQ 1.4 How can acoustic distance perception be improved?

RQ 2 What are viable non-visual multimodal interaction techniques?

RQ 2.1 What are the advantages and disadvantages of different tactile interaction techniques?

RQ 3 What is a good way to help users obtain an overview of available items and options?

RQ 4 Which interface metaphors fit the design space and comply with the WIMP paradigm?

RQ 5 How can the focus of attention be supported?

The results of the evaluation lead to a holistic design concept for eyes-free interfaces described in chapter 8.

Chapter 2 addressed RQ 1, 2, 3, 4, and 5. This chapter focused on introductions to important themes and research methods from the disciplines of psychoacoustics (such as spatial hearing, auditory memory, masking effects, and distance perception), psychology (such as attention and distraction, cognitive load, and the working memory), and presence research. The chapter provided an introduction to specific subject areas that play a key role in certain research aspects presented in later chapters. Chapter 2 proceeds to define and discuss auditory display components, such as auditory icons and earcons, the application of spatiality, and the simultaneous presentation of sounds. An in-depth review of the concept of cognitive metaphors and their application in interface design revealed the importance of a consistent and adequate use of metaphors. The Chapter finished with an inventory of

auditory displays and applications using sound as their main modality. One of the main conclusions of this chapter was that auditory user interfaces are a good alternative for use in predominantly mobile domains. Auditory interfaces are similar to tactile interfaces in that they are flexible, scalable, and, most importantly, they do not interfere with visual information processing.

Chapter 3 addressed RQ 1.3 and 1.4. This chapter explored a technique based on simplified echolocation, which was developed to help users perform spatial cognition tasks on mobile devices and to support the focus of attention in multitasking environments. The method works as follows: users send a request, objects in the environment (e.g. active applications in an operating system or landmarks in a navigation system) reply, revealing their distance from the user through the time the sound waves take to travel from the object to the user. After an introduction to prior applications of sonar techniques in auditory interfaces, a user study is presented comparing the above method with verbally coded distance information. The results show that both methods have their advantages: verbally coded information is significantly more accurate when precise knowledge about the position of an object is sought, while echolocation is advantageous when users want to know which out of several objects is closest to them. As distance is mapped to the speed of sound, the way echolocation was implemented in this research is not suitable for displaying distances beyond 3000 meters. In the echolocation condition the total duration of the playback depends not on the number of objects but on the distance of the farthest object from the user, making this method suitable for the representation of a large number of close objects while the efficiency of the verbal method is defined by the total number of objects and decreases significantly beyond six to eight objects.

Chapter 4 addressed RQ 1.1, 1.2, and 3. In this chapter the effects of certain design choices on the users' ability to detect one specific item among many and to gain an overview of items contained in a list is

studied. The design choices were: the type of sound samples, their combination, concatenation, and the playback setup. The chapter begins with an introduction, then proceeds to give an overview of prior related work on the intelligibility of simultaneously played sounds and important influential factors. Subsequently, the results of a user study were summarized. The findings showed that a significant decrease of error rates is yielded with each increase in interstimulus onset intervals (50, 100, 200 and 400 ms). There were no substantial differences between monophonic headphone and multichannel loudspeaker playback, nor between text-to-speech synthesised sound samples and earcons. The chapter concluded with a discussion of the results and recommendations for the presentation of overview information: Onset delays of 200 ms are a good trade-off between overall playback time and error rate. Non-spatialised headphone playback is sufficient if location information is not needed. Synthesized speech performs comparably to earcons and should be preferred because it is easy to produce, versatile and does not require prior memorization.

Chapter 5 addressed RQ 1. In this chapter an investigation was conducted into the effect of movement patterns in a spatial sound space on the perceived amount of simulator sickness, the pleasantness of the experience, and the perception of workload. After an introduction to the subject and a brief overview of related work on vection and psychoacoustic experiments on the perception of sound source movements, the setup and results of a user study are described. In the study the impact on symptoms of simulator sickness induced by regular movement patterns were compared to random patterns or no movements. Nearly 48 percent of all participants showed mild to moderate symptoms of simulator sickness, with a trend towards stronger symptoms for those experiencing left-right movements. Evidence was found for predictable left-right movements leading to a perceived unpleasantness that is significantly higher than for unpredictable or no movement at all. No noticeable effect on the perceived cognitive load for simple tasks was found for any of the conditions.

Chapter 6 addressed RQ 1.2. In this chapter the impact of monophonic, stereophonic, and binaural human speech recordings was studied in regard to their ability to induce the feeling of presence and influence the understanding of the speaker's emotional state. Speech based applications have long been core components of mobile phones and it seems reasonable to expect a broadening beyond the one-to-one phone call towards multiparty calls, speech based social networking and entertainment applications and text-to-speech renderings of written content. The chapter briefly touched on the concepts of presence and social presence (which are discussed more thoroughly in chapter 2) and gives an overview of prior related work on the perception of different sound recording techniques. The main part of the chapter is dedicated to a user study and a discussion of the derived results which show a significant advantage of binaural over mono and stereo sound for inducing the sense of being present in a virtual environment. It was found that listening to a stereophonic recording of a conversation leads to a significantly better understanding of the emotional state of speakers than when listening to a mono or binaural recording. Thus, if a sense of spatiality and presence is not required, stereophonic sound is a sufficient sound reproduction method for speech-based communication applications. Otherwise the use of binaural sound reproduction was recommended.

Chapter 7 addressed RQ 2, 2.1, and 5. This chapter comprises three individual user studies. After a general introduction to the subject of user input techniques these studies and their individual contexts were described.

The first study explored the design space of tangible interaction with a mobile auditory interface. The study design was geared towards delivering insights into which gestures users generate intuitively when interacting with a spatial sound space. After a description of the study setup, the results were discussed in terms of the scope of the gestures proposed, their tangible aspects, and the users' own preferences. Recommendations were derived for the design of gesture based interaction

techniques for multimodal displays, such as: Gestures should be minimally expressive and restricted to small, context related sets; inversive gestures should be supported; clear and distinctive feedback for successful or unsuccessful gestures is essential.

The second study was an empirical comparison of gesture-based and key-based interaction techniques using the example of an application supporting synchronous multiparty voice communication. Earlier ideas, such as using the notion of distance to assign focus and manage attention (chapter 3) and the use of ontological and orientational interface metaphors (chapter 2), were incorporated into the design of the exemplary interface. While the results showed that traditional keypad interaction proved to be more straightforward to use, there was no significant impact on task completion times or the number of interaction movements made between the techniques. Overall, users felt that the spatial audio application supported group awareness while aiding peripheral task monitoring. They also felt that spatial audio support aided the feeling of social connectedness and offered enhanced support for communication.

The last study presented in chapter 7 evaluated the impact of visual and auditory display techniques on multitask performance in the context of a driving simulation. A short introduction was given to the psychological concept of attention in vehicle related research[1], pointing out that tasks, which rely on the same sensory resource, compete for attention and result in a degraded performance in both tasks when attended to simultaneously. After a summary of the related work the study design was elaborated. In the study participants were asked to perform tasks of varying difficulty under one visual and two different auditory conditions. The auditory interfaces proved to be as fast as the visual, while at the same time providing a lesser distraction from the primary driving task. Results from a post-study questionnaire showed a higher user satisfaction with the auditory user interfaces and workload

[1] Which was an expansion of the more generic section on attention an distraction given in chapter 2.

measurements showed lower impact when using the auditory interfaces. After a discussion of the results this section concluded by giving design recommendations for in-vehicle multimodal displays: Auditory interfaces are a good alternative to visual interfaces. Verbal output is very effective for shorter tasks; good feedback on the current position in the menu should be provided; it should be refrained from the parallel playback of several sounds should be prevented; the scrolling device, which was attached to the steering wheel, is a very effective interaction device for both visual and the auditory interfaces.

9.2 Foogue – A Holistic Design Concept for Smartphones

The information gathered and the lessons learned in these previous chapters led to the design of *Foogue*, an eyes-free interface concept for smart mobile devices. As previous work and the work presented in chapter 7 has demonstrated the usefulness of auditory display techniques in mobile scenarios, *Foogue* refrains from using visual elements and hence does not compete for the user's limited amount of visual attention. Drawing from the lessons learned, a gesture language was designed that is built from a limited number of simple but also easy to differentiate gesture elements. Alternatively, *Foogue* can support keyboard input. However, due to the novelty of the approach, the gestural input technique was further elaborated and discussed, while the keyboard interaction technique was not further conceptualized.

Foogue offers a design solution that is adapted to the challenges of operating a small device in a mobile situation. Smartphone usage patterns deviate from desktop or laptop usage patterns in so far as the mobile user spends less time actively interacting with the device and more time passively consuming. *Foogue* addresses these different usage patterns by offering two modes: *Menu Mode* and *Listening Mode*. While *Menu Mode* is designed to grant quick access to files, *Listening Mode* makes listening to files as comfortable as possible.

In *Menu Mode* the file structure is presented in a 120 degree to 180 degree arc in front of the user. Spatialised sound objects represent folders and items. The user can scan the content of the current folder by moving the

phone like a torch along the sound sources; by doing so the item currently pointed at will be read to the user as a word, e.g. 'music' or 'contacts'. A prototype programmed to investigate the viability of this methods was developed and showed promising results in an informal evaluation. However, further investigations are required to study these selection procedures more closely.

By performing the *open* gesture on a container the user descends into the hierarchy. Applying the *open* gesture on a file will pass it on to a *player*. *Players* resemble *windows* in graphical user interfaces. Depending on the type of file, the *player* is either a text-to-speech engine, a music player, a phone call pipe, or an acoustic notifier for new events, etc. Once a *player* is initiated it is displayed in *Listening Mode*, which can be entered by performing the *switch mode* gesture. *Players* are spatialised and initially positioned in front of the user. Each player can be selected by pointing the phone at the desired *player*. Players can be repositioned on a 360 degree circle around the user with the *drag & drop* gesture. If multiple *players* are active, the user can either focus on a *player* by pulling it closer or by pushing other *players* away so they are playing from a distance and are accordingly reduced in volume. In this way *Foogue* supports multitasking but also offers an analogy to the 'minimize' and 'maximize' or 'foreground' and 'background' options in visual interfaces.

Foogue allows users to interact with their smartphone in mobile situations without competing for visual attention. The interface is optimized for mobile usage patterns and although it is designed to be self-contained and fully functional, it can be complemented with visual output or alternative interaction techniques. *Foogue* is a high-level interface and hence does not require particular hardware but works on state-of-the-art smartphones.

9.3 Contributions

The work presented in this thesis has demonstrated the potential of audio in eyes-free interfaces for mobile devices. In particular, spatial audio has been shown to have a wide range of benefits. By focussing on these benefits, eyes-free interfaces can be built that are similarly structured and as

efficient as generic graphical user interfaces but have fewer disadvantages in mobile situations. Tactile user input in the form of 2D and 3D gestures proved to be an effective and enjoyable way of interacting with an auditory interface. While in this thesis the strengths of audio were highlighted in this thesis, weaknesses were also addressed. The results gained in this process are manifold and applicable beyond the scope of this thesis. Both mobile social networking applications and stationary systems can exploit audio to increase the sense of presence in a shared space and the feeling of connectedness to others. GPS based navigation tools can deliver distance information not only verbally but with the support of the methods proposed in this thesis. The holistic design concept presented at the end of this thesis is only one way to translate the accumulated insights into a usable interface. Although this thesis was not primarily motivated by the need to enable blind and visually impaired users to interact with smart devices, the insights gained and design guidelines derived can contribute to the research and development of interfaces in that field.

9.4 Limitations and Future Work

Foogue is first and foremost a design concept, not a fully developed interface. Many of the core features as defined along the WIMP paradigm and set out in the introduction have been addressed:

- *Players* in *Listening Mode* resemble Windows and support:
  - grouping
  - content retrieval
  - a way to focus attention (by means of positioning)
- *Synthesized spoken name handlers* resemble Icons:
  - selectable sound sources positioned in 3D space present objects and containers as entities
- *Menu Mode* presents Menus and supports:
  - a way to access objects and containers
  - a way to gain an overview of structures, items, and options

- *Gestures* resemble a Pointing Device and support:
  - navigate through hierarchical structures
  - selection and manipulation of objects

However, many questions have been raised throughout the course of this thesis, which point the way for further avenues for research. While some interesting topics for future research have been outlined in previous chapters the following paragraphs give a more general overview:

2D & 3D Gestures: The explorative study presented in section 7.1 found that users intuitively used both 2D and 3D gestures as well as combinations of both. The following studies, however, only focused on 3D gestures and only a small subset of these were incorporated into *Foogue*. This is somewhat unsatisfactory as touch screen gestures have the potential to both supplement or substitute 3D gestures. The prototypes described in section 8.3.4 are a first step towards exploring the applicability and user friendliness of combinations between the two methods. A more thorough evaluation is necessary from which a complete set of 2D and 3D gestures can be derived.

Distance: A method for the display of distance using the travel time of sound has been shown to be very effective when users want to know which out of several objects is closest to them. The discussion has shown that the effectiveness of the method for other tasks is highly dependent on proper sound source localisation. For the presented prototype, only interaural time differences (ITD) as localisation cues were applied. It would be worth exploring how other localisation cues, such as interaural level differences (ILD) or echoic room simulations can improve the sound source localisation and hence the effectiveness of the method for other tasks. Furthermore, it would be interesting to explore ways of displaying distances beyond 2000 meters without ever longer display times.

Item Detection and Overview Information: In chapter 4 recommendations were given for the design of applications utilising audio to display

several objects contained in a scene. The rapid playback technique explored in the study may be a good candidate for providing overviews of folder contents in *Foogue*. The rapid playback could also be used in a way that is similar to task switching interfaces (pressing ALT+TAB on Windows or CMD+TAB on OS X) to give an overview of instantiated *players*. Further research is needed to identify the advantages that this method can provide in other application areas such as GPS based navigation software or gaming.

Social Networking: The trend towards social networking services with growing communities has been picked up by device producers and application developers. These networks stimulate, if not require, a continuous and constant care and attention, which is enabled through mobile devices that are "always on". While synchronous and asynchronous information exchange is rapidly increasing, the methods for displaying this information lags behind. It is reasonable to assume that better, more capable alternatives to the hitherto existing methods of information displays will emerge. These could be group oriented voice chat spaces similar to the Thunderwire application developed by Hindus et al. [193] or, the more recent client/server based voice chats like Teamspeak[2] and Ventrilo[3]. It seems reasonable to assume that once "hear-through" hardware like the Augmented Reality Audio Headset [9] depicted in figure 5.3.3 is ready for the market, mobile audio spaces for cooperative work, game play and social networking will thrive. A first interface designed to support multiparty chats has been proposed in section 7.2 of chapter 7, and an evaluation of the impact of sound reproduction methods on the perception of social presence was described in 6. Although the results were incorporated into *Foogue*, which is per se capable of supporting such speech based services, long-term studies are required to identify the central issues and derive requirements for the design of appropriate interfaces.

[2] `http://www.teamspeak.com/`

[3] `http://www.ventrilo.com/`

Using the ideas presented in this paper, one would hope that in the future someone strolling through the Tiergarten park in Berlin could participate in a text message interaction without being removed from the enjoyment of a fine day in pleasant surroundings.

References

[1] J. Sodnik, C. Dicke, and T. Saso, "Auditory interfaces for mobile devices," *Encyclopedia of Wireless and Mobile Communications*, no. 1, pp. 1–9, 2010.

[2] C. Dicke, V. Aaltonen, A. Rämö, and M. Vilermo, "Talk to me: The influence of audio quality on the perception of social presence.," in *Proceedings of the Conference on Human Computer Interaction (HCI'10)*, (Dundee, Scotland), 2010.

[3] C. Dicke, V. Aaltonen, and M. Billinghurst, "Occurrence of simulator sickness in spatial sound spaces and 3d auditory displays," in *Proceedings of the 15th International Conference on Auditory Display (ICAD'09)* (M. Aramaki, R. Kronland-Martinet, S. Ystad, and K. Jensen, eds.), (Copenhagen, Denmark), 2009.

[4] C. Dicke, V. Aaltonen, and M. Billinghurst, "Simulator sickness in mobile spatial sound spaces," in *Proceedings of the 6th International Symposium CMMR/ICAD 2009* (S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, eds.), pp. 287–305, Springer, 2010.

[5] K. Wolf, C. Dicke, and R. Grasset, "Touching the void: Gestures for auditory interfaces," in *Proceedings of the 5th International Conference on Tangible, Embedded, and Embodied Interaction (TEI'11)*, (Funchal, Portugal), pp. 305–308, 2011.

[6] C. Dicke, S. Deo, M. Billinghurst, and J. Lehikoinen, "Experiments in mobile spatial audio-conferencing: Key-based and gesture-based interaction," in *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'08)*, (Amsterdam, Netherlands), pp. 91–100, 2008.

[7] J. Sodnik, C. Dicke, S. Tomazie, and M. Billinghurst, "A user study of auditory versus visual interfaces for use while driving," *Int. J. Human-Computer Studies*, vol. 66, no. 5, pp. 318–332, 2008.

[8] C. Dicke, K. Wolf, and Y. Tal, "Foogue: Eyes-free interaction for smartphones," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'10)*, (Lisbon, Portugal), pp. 455–458, 2010.

[9] M. Tikander, M. Karjalainen, and V. Riikonen, "An augmented reality audio headset," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx'08)*, (Espoo, Finland), pp. 181–184, 2008.

[10] W. Wierwille and L. Tijerina, "Vision in vehicles vi," in *Modelling the Relationship Between Driver In-Vehicle Visual Demands and Accident Occurrence* (A. Gale, I. Brown, C. Haslegrave, and S. Taylor, eds.), pp. 233–244, Elsevier, 1998.

[11] M. Sodhi, J. Cohen, and S. Kirschenbaum, "Multi-modal vehicle display design and analysis," Tech. Rep. URITC FY99-04, University of Rhode Island Transportation Center, 2004.

[12] G. Lakoff and M. Johnson, *Metaphors We Live By.* Chicago, IL, USA: University of Chicago Press, 1980.

[13] E. D. Mynatt, M. Back, R. Want, M. Baer, and J. B. Ellis, "Designing audio aura," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'98)*, (Los Augeles, CA, USA), pp. 566–573, 1998.

[14] N. Sawhney and C. Schmandt, "Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments," *ACM Trans. Comput.-Hum. Interact.*, vol. 7, no. 3, pp. 353–383, 2000.

[15] J. Williamson, R. Murray-Smith, and S. Hughes, "Shoogle: Excitatory multimodal interaction on mobile devices," in *Proceedings of*

the SIGCHI conference on Human Factors in Computing Systems (CHI'07), (San Jose, CA, USA), pp. 121–124, 2007.

[16] A. Pirhonen, S. A. Brewster, and C. Holguin, "Gestural and audio metaphors as a means of control for mobile devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02)*, (Minneapolis, MN, USA), pp. 291–298, 2002.

[17] J. D. Cutnell and K. W. Johnson, *Physics*. New York, NY, USA: Wiley, 8th ed., 2009.

[18] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. Colchester, UK: Informa Healthcare, 5th edition ed., 2010.

[19] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual Review of Psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[20] D. R. Begault, *3-D Sound For Virtual Reality and Multimedia.* Cambridge, UK: Academic Press, 1994.

[21] D. W. Batteau, "The role of the pinna in human localization," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 168, no. 1011, pp. 158–180, 1967.

[22] M. B. Gardner, "Some monaural and binaural facets of median plane localization," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1489–1495, 1973.

[23] A. W. Mills, "Auditory localization," in *Foundations of Modern Auditory Theory* (J. V. Tobias, ed.), pp. 303–348, New York, NY, USA: Academic Press, 1972.

[24] C. L. Searle, L. D. Braida, M. F. Davis, and H. S. Colburn, "Model for auditory localization," *The Journal of the Acoustical Society of America*, vol. 60, no. 5, pp. 1164–1175, 1976.

[25] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940.

[26] I. J. Hirsch, "Masking of speech and auditory localization," *International Journal of Audiology*, vol. 10, no. 2, pp. 110–114, 1971.

[27] G. J. Thomas, "Experimental study of the influence of vision on sound localization," *Journal of Experimental Psychology*, vol. 28, no. 2, pp. 163–177, 1941.

[28] D. A. Burgess, "Techniques for low cost spatial audio," in *Proceedings of the 5th Annual ACM Symposium on User Interface Software and Technology (UIST'92)*, (Monteray, CA, USA), pp. 53–59, 1992.

[29] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[30] W. A. Yost, R. H. Dye, and S. Sheft, "A simulated cocktail party with up to three sound sources," *Perception and Psychophysics*, vol. 58, pp. 1026–1036, 1996.

[31] D. R. Perrott and J. Tucker, "Minimum audible movement angle as a function of signal frequency and the velocity of the source," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1522–1527, 1988.

[32] D. W. Grantham, "Detection and discrimination of simulated motion of auditory targets in the horizontal plane," *The Journal of the Acoustical Society of America*, vol. 79, no. 6, pp. 1939–1949, 1986.

[33] D. R. Perrott and S. Pacheco, "Minimum audible angle thresholds for broadband noise as a function of the delay between the onset of the lead and lag signals," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2669–2672, 1989.

[34] R. S. Woodworth, *Experimental Psychology.* New York, NY, USA: Holt, 1938.

[35] L. Demany and C. Semal, "The role of memory in auditory perception," in *Auditory Perception of Sound Sources* (W. A. Yost, A. N. Popper, and R. R. Fay, eds.), vol. 29, pp. 77–113, New York, NY, USA: Springer, 2008.

[36] W. A. Yost, *Fundamentals of Hearing: An Introduction.* San Diego, CA, USA: Academic Press, Elsevier, 5th ed., 2006.

[37] P. D. Coleman, "An analysis of cues to auditory depth perception in free space," *Psychological Bulletin*, vol. 60, pp. 302–315, 1963.

[38] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[39] P. Cochran, J. Throop, and W. E. Simpson, "Estimation of distance of a source of sound," *American Journal of Psychology*, vol. 81, no. 2, pp. 198–206, 1968.

[40] P. Zahorik, "Auditory display of sound source distance," in *Proceedings of the International Conference on Auditory Display (ICAD'02)*, (Kyoto, Japan), 2002.

[41] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[42] D. H. Mershon, "Phenomenal geometry and the measurement of perceived auditory distance," in *Binaural and Spatial Hearing in Real and Virtual Environments* (R. H. Gilkey and T. R. Anderson, eds.), pp. 257–274, Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1997.

[43] P. D. Coleman, "Dual role of frequency spectrum in determination of auditory distance," *The Journal of the Acoustical Society of America*, vol. 44, no. 2, pp. 631–632, 1968.

[44] D. R. Perrott, T. R. Sadralodabai, K. Saberi, and T. Z. Strybel, "Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 33, pp. 389–400, 1991.

[45] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, Il, V. Balasubramanian, and P. Sterling, "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428–1434, 2006.

[46] H. Jacobson, "The informational capacity of the human eye," *Science*, vol. 113, no. 2933, pp. 292–293, 1951.

[47] B. Shinn-Cunningham, H. Lehnert, G. Kramer, E. Wenzel, and N. Durlach, "Auditory displays," in *Binaural and Spatial Hearing in Real and Virtual Environments* (R. H. Gilkey and T. Anderson, eds.), pp. 611–663, New York, NY, USA: Erlbaum, 1997.

[48] M. A. Nees and B. N. Walker, "Auditory interfaces and sonification," in *The Universal Access Handbook* (C. Stephanidis, ed.), pp. 507–522, L. Erlbaum Associates, New York, 2009.

[49] W. James, *The Principle of Psychology.* New York, NY, USA: Holt, 1890.

[50] D. Navon and D. Gopher, "On the economy of the human processing system," *Psychological Review*, vol. 86, no. 3, pp. 214–255, 1979.

[51] C. D. Wickens, "Processing resources in attention," in *Varieties of Attention* (R. Parasuraman and R. Davies, eds.), pp. 63–102, New York, NY, USA: Academy Press, 1984.

[52] P. Green, "Crashes induced by driver information systems and what can be done to reduce them," in *In Proceedings of Convergence 2000, Soc. of Automotive Engineers*, (Warrendale, PA, USA), pp. 26–36, Society of Automotive Engineers, 2000.

[53] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159–177, 2002.

[54] J. D. Lee, B. Caven, S. Haake, and T. L. Brown, "Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, no. 4, pp. 631–640, 2001.

[55] A. J. McKnight and A. S. McKnight, "The effect of cellular phone use upon driver attention," *Accident Analysis & Prevention*, vol. 25, no. 3, pp. 259–265, 1993.

[56] T. A. Ranney, E. Mazzae, R. Garrott, and M. J. Goodman, "Nhtsa driver distraction research: Past, present, and future," 2000.

[57] S. T. Iqbal, Y.-C. Ju, and E. Horvitz, "Cars, calls, and cognition: Investigating driving and divided attention," in *Proceedings of the 28th International Conference on Human factors in Computing Systems (CHI'10)*, (Atlanta, GA, USA), pp. 1281–1290, 2010.

[58] K. L. Young, M. A. Regan, and M. Hammer, "Driver distraction: A review of the literature," Tech. Rep. 206, Monash University Accident Research Centre, Victoria, Australia, 2003.

[59] M. Vollrath and I. Trotzke, *In-Vehicle Communication and Driving: An Attempt to Overcome their Interferences.* Würzburg, Germany: Center for Traffic Sciences, IZVW, University of Wuerzburg, 2000.

[60] E. Spelke, I. J. Hirsch, and U. Neisser, "Skills of divided attention," *Cognition*, vol. 4, no. 3, pp. 215–230, 1976.

[61] W. Hirst, E. S. Spelke, C. C. Reaves, G. Caharack, and U. Neisser, "Dividing attention without alternation or automaticity," *Journal of Experimental Psychology: General*, vol. 109, no. 1, pp. 98–117, 1980.

[62] W. Hirst and D. Kalmar, "Characterizing attentional resources," *Journal of Experimental Psychology: General*, vol. 116, no. 1, pp. 68–81, 1987.

[63] E. R. Hafter, A. Sarampalis, and P. Loui, "Auditory attention and filters," in *Auditory Perception of Sound Sources* (W. A. Yost, A. N. Popper, and R. R. Fay, eds.), vol. 29 of *Springer Handbook of Auditory Research*, pp. 115–142, Springer US, 2007.

[64] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[65] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12 (July), pp. 35–50, 1992.

[66] L. J. Stifelman, "The cocktail party effect in auditory interfaces: A study of simultaneous presentation," tech. rep., MIT Media Laboratory, 1994.

[67] J. Cohen, "Monitoring background activities," in *Proceedings of the International Conference on Auditory Display (ICAD'92)* (G. Kramer, ed.), (Santa Fé, NM, USA), pp. 499–532, Addison-Wesley, 1992.

[68] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker condition," *Acoustica*, vol. 86, pp. 117–128, 2000.

[69] D. E. Broadbent, *Perception and Communication*. London, UK: Pergamon, 1958.

[70] A. M. Treisman, "Contextual cues in selective listening," *Quarterly Journal of Experimental Psychology*, vol. 12, no. 4, pp. 242–248, 1960.

[71] N. Moray, "Broadbent's filter theory - postulate h and the problem of switching time," *The Quarterly Journal of Experimental Psychology*, vol. 12, pp. 214–220, 1960.

[72] A. M. Treisman, "Selective attention in man," *British Medical Bulletin*, vol. 20, no. 1, pp. 12–16, 1964.

[73] A. M. Treisman, "Strategies and models of selective attention," *Psychological Review*, vol. 76, no. 3, pp. 282–299, 1969.

[74] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," *Psychological Review*, vol. 70, pp. 80–90, 1963.

[75] D. A. Norman, "Toward a theory of memory and attention," *Psychological Review*, vol. 75, no. 6, pp. 522–536, 1968.

[76] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti, "Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'05)*, (Portland, OR, USA), pp. 919–928, 2005.

[77] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[78] A. Baddeley and G. Hitch, "Working memory," in *The Psychology of Learning and Motivation: Advances in Research and Theory* (G. H. Bower, ed.), vol. 8, pp. 47–89, New York, NY, USA: Academic Press, 1974.

[79] A. Baddley, *Working Memory*. New York, NY, USA: Oxford University Press, 1986.

[80] A. Baddeley, "Working memory: Looking back and looking forward," *Nature Reviews Neuroscience*, vol. 4, no. 10, pp. 829–839, 2003.

[81] A. D. Baddeley, N. Thomson, and M. Buchanan, "Word length and the structure of short-term memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 14, no. 6, pp. 575–589, 1975.

[82] A. Baddeley, "The episodic buffer: A new component of working memory?," *Trends in Cognitive Sciences*, vol. 4, no. 11, pp. 417–423, 2000.

[83] S. Oviatt, "Human-centered design meets cognitive load theory: Designing interfaces that help people think," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM'06)*, (Santa Barbara, CA, USA), pp. 871–880, 2006.

[84] S. Hart and L. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload* (P. Hancock and N. Meshkati, eds.), pp. 139–183, 1988.

[85] G. B. Reid, S. S. Potter, and J. R. Bressler, "Subjective workload assessment technique (swat): A user's guide," tech. rep., Armstrong Aerospace Medical Research Laboratory, 1989.

[86] J. H. McCracken and T. B. Aldrich, "Analysis of selected lhx mission functions: Implications for operator workload and system automation goals," tech. rep., Army Research Institute Aviation Research and Development Activity, 1984.

[87] J. Kjeldskov and C. Graham, "A review of mobile hci research methods," *Human-Computer Interaction with Mobile Devices and Services*, vol. 2795, no. Lecture Notes in Computer Science (LNCS), pp. 317–335, 2003.

[88] J. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. New York, NY, USA: John Wiley and Sons, 1976.

[89] F. Biocca and K. Nowak, "Plugging your body into the telecommunication system: Mediated embodiment, media interfaces, and social virtual environments," in *Communication Technology and Society* (C. Lin and D. Atkin, eds.), pp. 407–447, Waverly Hill, VI: Hampton Press, 2001.

[90] C. Heeter, "Being there: The subjective experience of presence," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 2, pp. 262–271, 1992.

[91] H. P. de Greef and W. A. IJsselsteijn, "Social presence in the photoshare tele-application," in *Proceedings of the 3rd International Workshop on Presence*, (Delft, Netherlands), 2000.

[92] M. Lombard and T. Ditton, "At the heart of it all: The concept of presence," *Journal of Computer-Mediated Communication*, vol. 3, no. 2, 1997.

[93] F. Biocca and C. Harms, "Defining and measuring social presence: Contribution to the networked minds theory and measure," in *Fifth Annual Workshop: Presence 2002*, (Universidade Fernando Pessoa, Porto, Portugal), pp. 7–36, 2002.

[94] S. Zhao, "Toward a taxonomy of copresence," *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 5, pp. 445–455, 2003.

[95] J. M. Loomis, "Distal attribution and presence," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 1, pp. 113–119, 1992.

[96] C. Hendrix and W. Barfield, "The sense of presence within auditory virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 3, pp. 290–301, 1996.

[97] A. W. Ellis and G. Beattie, *The Psychology of Language and Communication*. London, UK: The Guilford Press, 1st ed., 1986.

[98] R. L. Daft and R. H. Lengel, "Information richness: A new approach to managerial behavior and organizational design," in *Research in Organizational Behavior* (L. L. Cummings and B. M. Staw, eds.), vol. 6, pp. 191–233, Greenwich, CT, USA: JAI Press, 1984.

[99] R. E. Rice, "Media appropriateness," *Human Communication Research*, vol. 19, no. 4, pp. 451–484, 1993.

[100] R. E. Rice, "The internet and health communication: A framework of experiences," in *The Internet and Health Communication. Experiences and Expectations* (R. E. Rice and J. E. Katz, eds.), pp. 5–46, Thousand Oaks, CA, USA: SAGE Publications, 2001.

[101] L. Trevino, R. H. Lengel, and R. L. Daft, "Media symbolism, media richness, and media choice in organizations," *Communication Research*, vol. 14, no. 5, pp. 553–574, 1987.

[102] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Urbana, IL, USA: University of Illinois Press, 1957.

[103] F. Biocca, C. Harms, and J. Burgoon, "Criteria and scope conditions for a theory and measure of social presence," in *Proceedings of the 4th International Workshop on Presence*, (Philadelphia, PA, USA), 2001.

[104] M. Slater, "How colorful was your day? why questionnaires cannot assess presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 4, pp. 484–493, 2004.

[105] B. E. Insko, "Measuring presence: Subjective, behavioural and physiological methods," in *Being There: Concepts, Effects and Measurements of User Presence in Synthetic Environments* (G. Riva, F. Davide, and W. A. IJsselsteijn, eds.), pp. 110–118, Amsterdam, Netherlands: IOS Press, 2003.

[106] J. Freeman, S. E. Avons, D. E. Pearson, and W. A. Ijsselsteijn, "Effects of sensory information and prior experience on direct subjective ratings

of presence," *Presence: Teleoperators and Virtual Environments*, vol. 8, no. 1, pp. 1–13, 1999.

[107] W. A. Ijsselsteijn, H. de Ridder, J. Freeman, and S. E. Avons, "Presence: Concept, determinants and measurement," *Proceedings of the SPIE, Human Vision and Electronic Imaging V*, vol. 3959, pp. 520–529, 2000.

[108] F. Biocca, C. Harms, and J. K. Burgoon, "Toward a more robust theory and measure of social presence: Review and suggested criteria," *Presence: Teleoperators and Virtual Environments*, vol. 12, no. 5, pp. 456–480, 2003.

[109] B. Reeves, B. Detenber, and J. Steuer, "New televisions: The effects of big pictures and big sound on viewer responses to the screen," in *Paper Presented to the Information Systems Division of the International Communication Association*, (Washington, DC, USA), 1993.

[110] N. Yankelovich, J. Kaplan, J. Provino, M. Wessler, and J. M. DiMicco, "Improving audio conferencing: Are two ears better than one?," in *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'06)*, (Banff, AB, Canada), pp. 333–342, 2006.

[111] J. J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, (Seattle, WA, United States), pp. 166–173, 2001.

[112] M. Droumeva, "Understanding immersive audio: Historical and sociocultural aspects of auditory displays," in *Proceedings of the International Conference on Auditory Display (ICAD'05)*, (Limerick, Ireland), 2005.

[113] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," vol. 3, no. 3, pp. 203–220, 1993.

[114] E. M. Kolasinski, "Simulator sickness in virtual environments," Tech. Rep. 1027, U.S. Army Research Institue for the Behavioral and Social Sciences, 1995.

[115] J. T. Reason and J. J. Brand, *Motion Sickness*. London, UK and New York, NY, USA: Academic Press, 1975.

[116] M. Treisman, "Motion sickness: An evolutionary hypothesis," *Science*, vol. 197, no. 4302, pp. 493–495, 1977.

[117] L. J. Hettinger and G. E. Riccio, "Visually induced motion sickness in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 3, pp. 306–310, 1992.

[118] M. E. McCauley and T. J. Sharkey, "Cybersickness: Perception of self-motion in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 3, pp. 311–318, 1992.

[119] B. E. Riecke, J. Schulte-Pelkum, M. N. Avraamides, M. Von Der Heyde, and H. H. Bülthoff, "Cognitive factors can influence self-motion perception (vection) in virtual reality," *ACM Transactions on Applied Perception*, vol. 3, no. 3, pp. 194–216, 2006.

[120] B. B. Bederson, "Audio augmented reality: A prototype automated tour guide," in *Conference Companion on Human Factors in Computing Systems (CHI'95)*, (Denver, CO, USA), pp. 210–211, 1995.

[121] J. Rozier, K. Karahalios, and J. Donath, "Hear&there: An augmented reality system of linked audio," in *Proceedings of the International Conference on Auditory Display (ICAD'00)*, (Atlanta, GA, USA), 2000.

[122] K. Lyons, M. Gandy, and T. Starner, "Guided by voices: An audio augmented reality system," in *Proceedings of the International Conference on Auditory Display (ICAD'00)*, (Atlanta, GA, USA), 2000.

[123] V. Sundareswaran, K. Wang, S. Chen, R. Behringer, J. McGee, C. Tam, and P. Zahorik, "3d audio augmented reality: Implementation and experiments," in *Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR'03)*, (Washington, DC, USA), pp. 296–297, 2003.

[124] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[125] L. Terrenghi and A. Zimmermann, "Tailored audio augmented environments for museums," in *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI'04)*, (Funchal, Madeira, Portugal), pp. 334–336, 2004.

[126] C. Stahl, "The roaring navigator: A group guide for the zoo with shared auditory landmark display," in *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'07)*, (Singapore), pp. 383–386, 2007.

[127] F. Heller, T. Knott, M. Weiss, and J. Borchers, "Multi-user interaction in virtual audio spaces," in *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI'09)*, (Boston, MA, USA), pp. 4489–4494, 2009.

[128] W. Gaver, "Auditory icons: Using sound in computer interfaces," *Human Computer Interaction*, vol. 2, no. 2, pp. 67–94, 1986.

[129] W. W. Gaver, "The sonic finder: An interface that uses auditory icons," *Human-Computer Interaction*, vol. 4, no. 1, pp. 67–94, 1989.

[130] S. A. Brewster, P. C. Wright, and A. D. Edwards, "An evaluation of earcons for use in auditory human-computer interfaces," in *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (S. Ashlund, A. Henderson, E. Hollnagel, K. Mullet,

and T. White, eds.), CHI '93, (New York, NY, USA), pp. 222–227, ACM, 1993.

[131] S. A. Brewster, "Navigating telephone-based interfaces with earcons," in *BCS HCI*, (Bristol, UK), pp. 39–56, 1997.

[132] S. A. Brewster, "Using non-speech sounds to provide navigation cues," *ACM Transactions on Computer-Human Interaction (TOCHI'98)*, vol. 5, no. 3, pp. 224–259, 1998.

[133] G. LePlâtre and S. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," in *Proceedings of the 6th International Conference on Auditory Display (ICAD'00)* (P. R. Cook, ed.), (Atlanta, GA, USA), pp. 190–199, 1998.

[134] S. A. Brewster and M. G. Crease, "Correcting menu usability problems with sound," *Behaviour & Information Technology*, vol. 18, no. 3, pp. 165–177, 1999.

[135] M. L. M. Vargas and S. Anderson, "Combining speech and earcons to assist menu navigation," in *Proceedings of the 9th International Conference on Auditory Display (ICAD'03)*, (Boston, MA, USA), 2003.

[136] S. D. Jones and S. M. Furner, "The construction of audio icons and information cues for human-computer dialogues," in *Contemporary Ergonomics: Proceedings of the Ergonomics Society's 1989 Annual Conference* (T. Megaw, ed.), (Reading, UK), pp. 436–441, Taylor & Francis, 1989.

[137] S. A. Brewster, *Providing a Structured Method for Integrating Non-Speech Audio into Human-Computer Interfaces*. Phd thesis, University of York, UK, 1994.

[138] P. Lucas, "An evaluation of the communicative ability of auditory icons and earcons," in *Proceedings of the 2nd International Conference on Auditory Display (ICAD'94)*, (Santa Fe, NM, USA), pp. 121–128, 1994.

[139] B. N. Walker, A. Nance, and J. Lindsay, "Spearcons: Speech-based earcons improve navigation performance in auditory menus," in *Proceedings of the 12th International Conference on Auditory Display (ICAD'06)*, (London, UK), pp. 63–68, 2006.

[140] L. Bölke and P. Gorny, "Direkte manipulation von akustischen objekten durch blinde rechnerbenutzer," in *Software-Ergonomie 95 – Fachtagung der German Chapter of the ACM und der Gesellschaft für Informatik.* (H.-D. Böcker, ed.), (Darmstadt, Germany), pp. 93–105, B.G. Teubner, 1995.

[141] P. Klante, "Praxisbericht zur gestaltung auditiver benutzeroberflächen.," in *1st annual GC-UPA Track*, (Stuttgart, Germany), pp. 57–62, 2003.

[142] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Workshop on the Applications of Signal Processing to Audio and Acoustics*, (New Platz, NY, USA), pp. 99–102, IEEE, 2001.

[143] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," *Journal of the AES*, vol. 49, pp. 231–249, 2001.

[144] E. Wenzel, M. Arruda, D. Kistler, and S. Foster, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustic Society of America*, vol. 94, pp. 111–123, 1993.

[145] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[146] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[147] V. Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, 2001.

[148] J. Jakka, *Binaural to Multichannel Audio Upmix*. Master's thesis, Helsinki University of Technology, Finland, 2005.

[149] T. Lossius, P. Baltazar, and T. de la Hogue, "Dbap - distance-based amplitude panning," in *Proceedings of the International Computer Music Conference (ICMC'09)*, (Montreal, QC, Canada), 2009.

[150] D. Kostadinov, J. D. Reiss, and V. Mladenov, "Evaluating distance-based amplitude panning for spatial audio," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, (Dallas, TX, USA), 2010.

[151] D. K. McGookin and S. A. Brewster, "Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition," *ACM Transactions on Applied Perceptions*, vol. 1, no. 2, pp. 130–155, 2004.

[152] D. S. Brungart, M. A. Ericson, and B. D. Simpson, "Design considerations for improving the effectiveness of multitalker speech displays," in *Proceedings of the International Conference on Auditory Display (ICAD'02)*, (Kyoto, Japan), 2002.

[153] D. S. Brungart and B. D. Simpson, "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal," *Journal of the Acoustic Society of America*, vol. 112, no. 2, pp. 664–676, 2002.

[154] J. Kildal, *Developing an Interactive Overview for Non-Visual Exploration of Tabular Numerical Information. PhD Thesis.* PhD thesis, University of Glasgow, Glasgow, UK, 2009.

[155] P. Barr, R. Biddle, and . Noble, "A taxonomy of user-interface metaphors," in *Proceedings of the SIGCHI-NZ Symposium On*

Computer-Human Interaction (CHINZ'02), (Hamilton, New Zealand), p. 13, School of Mathematical and Computing Sciences, Victoria University of Wellington, 2002.

[156] J. Hurtienne and J. H. Israel, "Image schemas and their metaphorical extensions: Intuitive patterns for tangible interaction," in *Proceedings of the 1st International Conference on Tangible and Embedded Interaction (TEI'07)*, (Baton Rouge, LA, USA), pp. 127–134, 2007.

[157] G. Lakoff, *Women, Fire, and Dangerous Things*. Chicago, IL, USA: University of Chicago Press, 1990.

[158] T. Krzeskowski, "The axiological parameter in preconceptual image schemata," in *Conceptualizations and Mental Processing in Language* (R. Geiger and B. Rudzka-Ostyn, eds.), vol. 3 of *Cognitive Linguistics Research*, pp. 307–329, Berlin, Germany: Mouton de Gruyter, 1993.

[159] F. Boers, *Spatial Prepositions and Metaphor: A Cognitive Semantic Journey Along the Up-Down and the Front-Back Dimensions*. Tübingen, Germany: G. Narr, 1996.

[160] G. Lakoff and M. Turner, *More than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago, IL, USA: University of Chicago Press, 1989.

[161] P. Barr, R. Biddle, and J. Noble, "A semiotic model of user-interface metaphor," in *Virtual, Distributed and Flexible Organisations. Studies in Organisational Semiotics* (K. Liu, ed.), pp. 189–215, Springer, 2005.

[162] M. Cohen and L. F. Ludwig, "Multidimensional audio window management," *International Journal of Man-Machine Studies*, vol. 34, no. 3, pp. 319–336, 1991.

[163] W. W. Gaver, R. B. Smith, and T. O'Shea, "Effective sounds in complex systems: The arkola simulation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'91)*, (New Orleans, LA, USA), pp. 85–90, 1991.

[164] P. M. Aoki, M. Romaine, M. H. Szymanski, J. D. Thornton, D. Wilson, and A. Woodruff, "The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations," in *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'03)*, (Ft. Lauderdale, FL, USA), pp. 425–432, 2003.

[165] W. Dell, "The use of 3d audio to improve auditory cues in aircraft," Tech. Rep. CS4H, Department of Computing Science, University of Glasgow, 1999.

[166] S. Goose and S. Djennane, "Wire3: Driving around the information super-highway," *Personal and Ubiquitous Computing*, vol. 6, pp. 164–175, 2002.

[167] S. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker, "Multimodal 'eyes-free' interaction techniques for wearable devices," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'03)*, vol. 5, (Ft. Lauderdale, FL, USA), pp. 473–480, 2003.

[168] M. Kobayashi and C. Schmandt, "Dynamic soundscape: Mapping time to space for audio browsing," in *Extended Abstracts on Human Factors in Computing Systems (CHI'97)*, (Atlanta, GA, USA), pp. 194–201, 1997.

[169] C. Frauenberger and T. Stockman, "Patterns in auditory menu design," in *Proceedings of the International Conference on Auditory Display (ICAD'06)*, (London, UK), pp. 141–147, 2006.

[170] K. Crispien, K. Fellbaum, A. Savidis, and C. Stephanidis, "A 3d-auditory environment for hierarchical navigation in non-visual interaction," in *Proceedings of the International Conference on Audio Display (ICAD'96)*, (Palo Alto, CA, USA), pp. 18–21, 1996.

[171] C. Schmandt, "Audiohallway: A virtual acoustic environment for browsing," in *Proceedings of the 11th Annual ACM Symposium on*

User Interface Software and Technology (UIST'98), (San Francisco, CA, USA), pp. 163–170, 1998.

[172] E. Costanza, J. Panchard, G. Zufferey, J. Nembrini, J. Freudiger, J. Huang, and J.-P. Hubaux, "Sensortune: A mobile auditory interface for diy wireless sensor networks," in *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10)*, (Atlanta, GA, USA), pp. 2317–2326, 2010.

[173] L. J. Stifelman, B. Arons, C. Schmandt, and E. A. Hulteen, "Voicenotes: A speech interface for a hand-held voice notetaker," in *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, (Amsterdam, Netherlands), pp. 179–186, 1993.

[174] A. Walker, S. A. Brewster, D. McGookin, and A. Ng, "Diary in the sky: A spatial audio display for a mobile calendar," in *Proceedings of BCS IHM-HCI*, (Lille, France), pp. 531–540, Springer, 2001.

[175] A. Walker and S. A. Brewster, "Spatial audio in small display screen devices," *Personal Technologies*, vol. 4, no. 2, pp. 144–154, 2000.

[176] S. Zhao, P. Dragicevic, M. Chignell, R. Balakrishnan, and P. Baudisch, "Earpod: Eyes-free menu selection using touch input and reactive audio feedback," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*, (San Jose, CA, USA), pp. 1395–1404, 2007.

[177] K. A. Li, P. Baudisch, and K. Hinckley, "Blindsight: Eyes-free access to mobile phones," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'08)*, (Florence, Italy), pp. 1389–1398, 2008.

[178] A. Kan, G. Pope, C. Jin, and A. van Schaik, "Mobile spatial audio communication system," in *Proceedings of the International Conference on Auditory Display (ICAD'04)*, (Sydney, Australia), 2004.

[179] S. Holland, D. R. Morse, and H. Gedenryd, "Audiogps: Spatial audio in a minimal attention interface," *Personal and Ubiquitous Computing*, vol. 6, no. 4, pp. 253–259, 2002.

[180] N. Mariette, "A novel sound localization experiment for mobile audio augmented reality applications," in *Advances in Artificial Reality and Tele-Existence* (Z. Pan, A. Cheok, M. Haller, R. W. H. Lau, and H. Saito, eds.), vol. 4282 of *Lecture Notes in Computer Science (LNCS)*, pp. 132–142, Berlin and Heidelberg, Germany: Springer, 2006.

[181] D. McGookin, S. A. Brewster, and P. Priego, "Audio bubbles: Employing non-speech audio to support tourist wayfinding," in *Haptic and Audio Interaction Design* (M. E. Altinsoy, U. Jekosch, and S. A. Brewster, eds.), vol. 5763 of *Lecture Notes in Computer Science (LNCS)*, pp. 41–50, Berlin and Heidelberg, Germany: Springer, 2009.

[182] E. Isaacs, A. Walendowski, and D. Ranganthan, "Hubbub: A sound-enhanced mobile instant messenger that supports awareness and opportunistic interactions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02)*, (Minneapolis, MN, USA), pp. 179–186, 2002.

[183] B. B. Blasch, R. G. Long, and S. N. Griffin, "Results of a national survey of electronic travel aid use," *Journal of Visual Impairment and Blindness*, vol. 33, no. 9, pp. 449–453, 1989.

[184] P. B. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[185] C. T. Davies, C. M. Burns, and S. D. Pinder, "Using ecological interface design to develop an auditory interface for visually impaired travellers," in *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments (OZCHI'06)*, (Sydney, Australia), pp. 309–312, 2006.

[186] G. Ghiani, B. Leporini, and F. Paternò, "Supporting orientation for blind people using museum guides," in *Extended Abstracts on Human Factors in Computing Systems (CHI'08)*, (Florence, Italy), pp. 3417–3422, 2008.

[187] V. Gaudissart, S. Ferreira, C. Thillou, and B. Gosselin, "Sypole: Mobile reading assistant for blind people," in *Proceedings of the 9th International Conference Speech and Computer (SPECOM'04)*, (St. Petersburg, Russia), pp. 538—-544, 2004.

[188] R. A. Kajastila and T. Lokki, "A gesture-based and eyes-free control method for mobile devices," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI EA'09)*, (Boston, MA, USA), pp. 3559–3564, 2009.

[189] I. Alvarez, M. Aqueasha, J. Dunbar, J. Taiber, D.-M. Wilson, and J. E. Gilbert, "Voice interfaced vehicle user help," in *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'10)*, (Pittsburgh, PA, USA), pp. 42–49, 2010.

[190] D. Svanaes and W. Verplank, "In search of metaphors for tangible user intefaces," in *Proceedings of DARE 2000 on Designing Augmented Reality Environments*, (Elsinore, Denmark), pp. 121–129, 2000.

[191] J. Cohen, "Out to lunch: Further adventures monitoring background activity," in *Proceedings of the International Conference on Auditory Display (ICAD'94)*, (Santa Fe, NM, USA), pp. 15–20, 1994.

[192] I. Smith and S. E. Hudson, "Low disturbance audio for awareness and privacy in media space applications," in *Proceedings of the 3rd ACM International Conference on Multimedia (MM'95)*, (San Francisco, CA, USA), pp. 91–97, 1995.

[193] D. Hindus, M. S. Ackerman, S. Mainwaring, and B. Starr, "Thunderwire: A field study of an audio-only media space," in *Computer*

Supported Cooperative Work (CSCW'96), (Cambridge, MA, USA), pp. 238–247, 1996.

[194] R. Srinivasan and P. P. Jovanis, "Effect of in-vehicle route guidance systems on driver workload and choice of vehicle speed: Findings from a driving simulator experiment," in *Ergonomics and Safety of Intelligent Driver Interfaces*, pp. 97–114, Hillsdale, NJ, USA: L. Erlbaum Associates, 1997.

[195] B. S. Jensen, M. B. Skov, and N. Thiruravichandran, "Studying driver attention and behaviour for three configurations of gps navigation in real traffic driving," in *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10)*, (Atlanta, GA, USA), pp. 1271–1280, 2010.

[196] T. Horberry, J. Anderson, M. A. Regan, T. J. Triggs, and J. Brown, "Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance," *Accident Analysis and Prevention*, vol. 38, no. 1, pp. 185–91, 2006.

[197] D. S. Brungart, "Speech-based distance cueing in virtual auditory displays," in *Human Factors and Ergonomics Society Annual Meeting Proceedings (IEA/HFES'00)*, vol. 44, pp. 714–717, 2000.

[198] M. Supa, M. Cotzin, and K. M. Dallenbach, "'facial vision': The perception of obstacles by the blind," *The American Journal of Psychology*, vol. 57, no. 2, pp. 133–183, 1944.

[199] T. A. Stroffregen and J. B. Pittenger, "Human echolocation as a basic form of perception and action," *Ecological Psychology*, vol. 7, no. 3, pp. 181–216, 1995.

[200] W. Schiff and R. Oldak, "Accuracy of judging time to arrival: Effects of modality, trajectory, and gender," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, no. 2, pp. 303–316, 1990.

[201] A. D. Heyes, "Human navigation by sound," *Physics in Technology*, vol. 14, no. 2, pp. 68–75, 1983.

[202] L. Kay, "Auditory perception of objects by blind persons, using a bioacoustic high resolution air sonar," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3266–3275, 2000.

[203] T. Shiose, K. Ito, and K. Mamada, "Identification of acoustic factors for perception of crossability common to blind and sighted pedestrians," in *Computers Helping People with Special Needs* (K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, eds.), vol. 4061 of *LNCS*, pp. 1273–1279, Berlin, Germany: Springer, 2006.

[204] M. Talbot and W. Cowan, "On the audio representation of distance for blind users," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*, (Boston, MA, USA), pp. 1839–1848, 2009.

[205] P. Baudisch and R. Rosenholtz, "Halo: A technique for visualizing off-screen objects," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*, (Ft. Lauderdale, FL, USA), pp. 481–488, 2003.

[206] M. Wertheimer, "Untersuchungen zur lehre von der gestalt ii," *Psychologische Forschung*, vol. 4, pp. 301–350, 1923.

[207] D. Cornish and D. Dukette, *The Essential 20: Twenty Components of an Excellent Health Care Team.* Pittsburgh, PA, USA: RoseDog Books, 2009.

[208] D. W. Massaro, "Preperceptual images, processing time, and perceptual units in auditory perception," *Psychological Review*, vol. 79, pp. 124–145, 1972.

[209] G. Lorho, J. Marila, and J. Hiipakka, "Feasability of multiple non-speech sounds presentation using headphones," in *Proceedings of the*

International Conference on Auditory Display (ICAD'01), (Espoo, Finland), 2001.

[210] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statistical Science*, vol. 18, no. 1, pp. 71–103, 2003.

[211] D. M. Johnson, "Introduction to and review of simulator sickness research," Tech. Rep. 1832, U.S. Army Research Institute - Rotary Wing Aviation Research Unit ATTN, Fort Rucker, AL, USA, 2005.

[212] R. S. Kennedy, L. J. Hettinger, and M. G. Lilienthal, "Simulator sickness," in *Motion and Space Sickness* (G. H. Crampton, ed.), pp. 317–341, Boca Raton, FL, USA: CRC Press, 1990.

[213] E. C. Regan and K. R. Price, "The frequency of occurrence and severity of side-effects of immersion virtual reality," *Aviation, Space, and Environmental Medicine*, vol. 65, no. 6, pp. 527–530, 1994.

[214] B. E. Riecke, D. Feuereissen, and J. J. Rieser, "Auditory self-motion illusions ('circular vection') can be facilitated by vibrations and the potential for actual motion," in *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization (APGV)* (S. H. Creem-Regehr and K. Myszkowski, eds.), (Los Angeles, CA, USA), pp. 147–154, ACM, 2008.

[215] T. Brandt, J. Dichgans, and E. Koeing, "Differential effects of central versus peripheral vision on egocentric and exocentric motion perception," *Experimental Brain Research*, vol. 16, no. 5, pp. 476–491, 1973.

[216] R. Pausch, T. Crea, and M. Conway, "A literature survey for virtual environments: Military flight simulator visual systems and simulator sickness," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 3, pp. 344–363, 1992.

[217] R. Dodge, "Thresholds of rotation," *Journal of Experimental Psychology*, vol. 6, no. 2, pp. 107–137, 1923.

[218] P. Larsson, D. Västfjäll, and M. Kleiner, "Perception of self-motion and presence in auditory virtual environments," in *Proceedings of the 7th Annual International Workshop of Presence*, (Valencia, Spain), pp. 252–258, 2004.

[219] S. Sakamoto, Y. Osada, Y. Suzuki, and J. Gyoba, "The effects of linearly moving sound images on self-motion perception," *Acoustical Science and Technology*, vol. 25, no. 1, pp. 100–102, 2004.

[220] A. Väljamäe, P. Larsson, D. Västfjäll, and M. Kleiner, "Auditory presence, individualized head-related transfer functions and illusory ego-motion in virtual environments," in *Proceedings of the 7th Annual International Workshop of Presence*, (Valencia, Spain), pp. 141–147, 2004.

[221] A. Väljamäe, "Auditorily-induced illusory self-motion: A review," *Brain Research Reviews*, vol. 61, no. 2, pp. 240–255, 2009.

[222] J. R. Lackner, "Induction of illusory self-rotation and nystagmus by a rotating sound-field," *Aviation, Space and Environmental Medicine*, vol. 48, no. 2, pp. 129–131, 1977.

[223] B. E. Riecke, A. Väljamäe, and J. Schulte-Pelkum, "Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality," *ACM Transactions on Applied Perception*, vol. 6, no. 2, pp. 1–27, 2009.

[224] N. R. Miller, N. J. Newman, V. Biousse, and J. B. Kerrison, *Walsh and Hoyt's Clinical Neuro-Ophthalmology*. Philadelphia, USA: Lippincott Williams & Wilkins, 6th ed., 2004.

[225] Y. A. Al'tman, O. V. Varyagina, V. S. Gurfinkel, and Y. S. Levik, "The effects of moving sound images on postural responses and the head

rotation illusion in humans," *Neuroscience and Behavioral Physiology*, vol. 35, no. 1, pp. 103–106, 2005.

[226] R. S. Kennedy, S. D. Lanham, C. J. Massey, and J. M. Drexler, "Gender differences in simulator sickness incidence: Implications for military virtual reality systems," *Safe Journal*, vol. 25, no. 1, pp. 69–76, 1995.

[227] F. Biocca, "Will simulation sickness slow down the diffusion of virtual environment technology?," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 3, pp. 334–343, 1992.

[228] D. S. Brungart and B. D. Simpson, "Improving multitalker speech communication with advanced audio displays," Tech. Rep. RTO-MP-HFM-123, Air Force Research Laboratory, AFRL/HECB, Wright-Patterson Air Force Base, OH, USA, 2005.

[229] M. J. Schuemie, P. van der Straaten, M. Krijn, and C. van der Mast, "Research on presence in virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 4, no. 2, pp. 183–201, 2001.

[230] M. Lombard and M. T. Jones, "Identifying the (tele)presence literature," *PsychNology Journal*, vol. 5, no. 2, pp. 197–206, 2007.

[231] W. Barfield, D. Zeltzer, T. Sheridan, and M. Slater, "Presence and performance within virtual environments," in *Virtual Environments and Advanced Interface Design*, pp. 473–513, Oxford University Press, 1995.

[232] J. Freeman and J. Lessiter, "Here, there and everywhere: The effects of multichannel audio on presence," in *Proceedings of the 7th International Conference on Auditory Display (ICAD'01)* (J. Hiipakka, N. Zacharov, and T. Takala, eds.), (Espoo, Finland), pp. 231–234, 2001.

[233] D. Västfjäll, "The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments," *Cyberpsychology & Behavior*, vol. 6, no. 2, pp. 181–188, 2003.

[234] V. Aaltonen, J. Takatalo, J. Häkkinen, M. Lehtonen, G. Nyman, and M. Schrader, "Measuring mediated communication experience," in *International Workshop on Quality of Multimedia Experience (QoMEx'09)*, (San Diego, CA, USA), pp. 104–109, 2009.

[235] M. Kylliäinen, H. Helimäki, N. Zacharov, and J. Cozens, "Compact high performance listening spaces.," in *Proceedings of Euronoise*, (Naples, Italy), 2003.

[236] C. DiStefano, M. Zhu, and D. Mîndrilă, "Understanding and using factor scores: Considerations for the applied researcher," *Practical Assessment, Research & Evaluation*, vol. 14, no. 20, 2009.

[237] O. Shaer and E. Hornecker, "Tangible user interfaces: Past, present, and future directions," *Found. Trends Hum.-Comput. Interact.*, vol. 3, no. 1-2, pp. 1–137, 2010.

[238] S. A. Brewster, R. Murray-Smith, A. Crossan, Y. Vasquez-Alvarez, and J. Rico, "The gaime project: gestural and auditory interactions for mobile environments," in *Whole Body Interaction Workshop*, (Boston, MA, USA), 2009.

[239] M. Karam and M. C. Schraefel, "A taxonomy of gestures in human computer interactions," Tech. Rep. ECSTR-IAM05-009, University of Southampton, Southampton, UK, 2005.

[240] J. Rico and S. A. Brewster, "Usable gestures for mobile interfaces: Evaluating social acceptability," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI'10)*, (Atlanta, GA, USA), pp. 887–896, 2010.

[241] S. Bhandari and Y.-K. Lim, "Exploring gestural mode of interaction with mobile phones," in *Extended Abstracts on Human Factors in Computing Systems (CHI'08)*, (Florence, Italy), pp. 2979–2984, 2008.

[242] C. S. Montero, J. Alexander, M. T. Marshall, and S. Subramanian, "Would you do that?: Understanding social acceptance of gestural interfaces," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'10)*, (Lisbon, Portugal), pp. 275–278, 2010.

[243] C. Lewis, "Using the 'thinking-aloud' method in cognitive interface design," Tech. Rep. IBM Research Report RC 9265, IBM Watson Research Center, 1982.

[244] K. Crispien and T. Ehrenberg, "Evaluation of the 'cocktail party effect' for multiple speech stimuli within a spatial audio display," *Journal of the Audio Engineering Society*, vol. 43, pp. 932–940, 1995.

[245] R. Drullman and A. W. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *Journal of the Accoustical Society of America*, vol. 107, no. 4, pp. 2224–2235, 2000.

[246] M. A. Ericson and R. L. McKinley, "The intelligibility of multiple talkers separated spatially in noise," in *Binaural and Spatial Hearing in Real and Virtual Environments* (R. H. Gilkey and T. R. Anderson, eds.), pp. 701–724, Mahwah, NJ, USA: Erlbaum, 1997.

[247] A. Savidis, C. Stephanidis, A. Korte, K. Crispien, and K. Fellbaum, "A generic direct-manipulation 3d-auditory environment for hierarchical navigation in non-visual interaction," in *Proceedings of the 2nd annual ACM conference on Assistive technologies (ASSETS'96)*, (Vancouver, BC, Canada), pp. 117–123, 1996.

[248] A. Walker and S. A. Brewster, "Extending the auditory display space in handheld computing devices," in *Proceedings of the 2nd Workshop on Human Computer Interaction with Mobile Devices*, (Edinburgh, UK), 1999.

[249] M. Billinghurst, J. Bowskill, and J. Morphett, "Wearcom: Wearable communication spaces," in *Proceedings of Collaborative Virtual Environments 1998 (CVE'98)*, (Manchester, UK), 1998.

[250] S. Goose, J. Riedlinger, and S. Kodlahalli, "Conferencing3: 3d audio conferencing and archiving services for handheld wireless devices," *International Journal of Wireless and Mobile Computing*, vol. 1, no. 1, pp. 5–13, 2005.

[251] S. Deo, M. Billinghurst, N. Adams, and J. Lehikoinen, "Experiments in spatial mobile audio-conferencing," in *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology*, (Singapore), pp. 447–451, 2007.

[252] J. F. Corso, "Age and sex differences in pure-tone thresholds: Survey of hearing levels from 18 to 65 years," *Arch Otolaryngol*, vol. 77, no. 4, pp. 385–405, 1963.

[253] J. D. Pearson, C. H. Morrell, S. Gordon-Salant, L. J. Brant, E. J. Metter, L. L. Klein, and J. L. Fozard, "Gender differences in a longitudinal study of age-associated hearing loss," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1196–1205, 1995.

[254] M. Wirth, H. Horn, T. Koenig, M. Stein, A. Federspiel, B. Meier, C. M. Michel, and W. Strik, "Sex differences in semantic processing: Event-related brain potentials distinguish between lower and higher order semantic analysis during word reading," *Cerebral Cortex*, vol. 17, no. 9, pp. 1987–1997, 2006.

[255] J. C. Stutts, D. W. Reinfurt, L. Staplin, and E. A. Rodgman, "The role of driver distraction in traffic crashes," tech. rep., University of North Carolina, Highway Safety Research Center, Chapel Hill, NC, USA, 2001.

[256] M. Goodman, F. Bents, L. Tijerina, W. Wierwille, N. Lerner, and D. Benel, "An investigation of the safety implications of wireless communications in vehicles," Tech. Rep. DOT-HS-808-635, US Department of Transportation, 1997.

[257] M. A. Pettitt, G. E. Burnett, and A. Stevens, "Defining driver distraction," in *Proceedings of the 12th ITS World Congress*, (San Francisco, CA, USA), 2005.

[258] L. Tijerina, "Issues in the evaluation of driver distraction associated with in-vehicle information and telecommunications systems," 2000.

[259] R. E. Llaneras, "Nhtsa driver distraction internet forum: Summary and proceedings," Final Report DTNH22-99-D-07005, Westat, 2000.

[260] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'88)*, pp. 213–218, 1988.

[261] G. Blaskoó and S. Feiner, "A menu interface for wearable computing," in *Proceedings of the 6th IEEE International Symposium on Wearable Computers (ISWC'02)*, (Seattle, WA, USA), pp. 164–165, 2002.

[262] G. N. Marentakis and S. A. Brewster, "Effects of feedback, mobility and index of difficulty on deictic spatial audio target acquisition in the horizontal plane," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*, (Montreal, QC, Canada), pp. 359–368, 2006.

[263] J. Raskin, *The Humane Interface: New Directions for Designing Interactive Systems*. New York, NY, USA: Addison-Wesley Publishing Co., 2000.

[264] Y. Perl and E. M. Reingold, "Understanding the complexity of interpolation search," *Information Processing Letters*, vol. 6, no. 6, pp. 219–222, 1977.